2460 IEEE Communications Magazine
28(1990)October, No.10, New York, US

State-Dependent Dynamic Traffic Management for Telephone Networks

Jean Regnier
W. Hugh Cameron

THE WIDESPREAD ADVENT OF COMPETITION IN providing telephone service has created strong incentive for suppliers to raise or preserve profitability by cutting costs and improving the completion rate of traffic offered to their networks. Dynamic traffic management is one means of satisfying this need. Applied in urban or intercity telephone networks, it has been shown to reduce trunking cost, reduce trunk network operating expenses, and reduce blocking of offered traffic under both normal and abnormal network stress conditions.

Although the basic principles of dynamic traffic routing for circuit-switched networks have been known for over twenty years [1], they have been applied only in the past decade. The introduction of stored program control switches with high-availability real-time data processing ability has made the application practical. Indeed, all currently implemented dynamic network traffic management systems (see companion articles) share the following characteristics:

- Switches apply different routing tables at different times for calls to a given destination, depending on network and traffic conditions.
- The changes in routing tables are based on current measurements of offered traffic and network performance, and are determined mechanically rather than by human calculation.
- Global network and traffic information is used (either directly or indirectly) to determine the routing tables applied locally by each switch.
- The routing tables permit mutual overflow among links at each switch so that, in principle, every link may carry calls overflowing from other links.

Thus, switches must be able to make frequent changes in their routing tables during periods of high calling; make and send out on-line measurements of calling and blocking rates for their traffic parcels; and apply different treatments to calls from different traffic parcels overflowing from a single link. They must also be capable of either making their own calculations to update their routing tables or accepting from outside the results of such calculations.

In the past decade, strides in microprocessing technology have given switches the administrative computing capacity to implement these new functions; hence, dynamic traffic management has become practical.

The distinguishing features of the system described in this article are:

- The explicit use of global network information, gathered from all participating switches, for each update of nodal routing tables and flow controls
- A very short update cycle (measurement, calculation, control)—typically 10 s—enabling the system together with the controlled network to operate as a closed feedback loop in real time

This system is called Dynamic Traffic Management (DTM). The traffic routing function of the system is being developed by Northern Telecom for the Canadian national and local telephone networks. Implementation is scheduled for 1991–1992. The congestion control function is still in the planning stage.

This article first presents a description of DTM. It then reviews the main design considerations which have led to the system's being designed as it is. Finally, it provides a quantitative overview of the benefits the system has been shown to provide in typical (Canadian) networks.

System Description

Overall System Architecture and Data Flow

Figure 1 depicts the architecture of DTM. The heart of the system is the network processor. It collects traffic operational measurements from the switches and in return makes recommendations regarding the traffic controls that they should enforce. Data collection, control selection and application are fully automated within a fixed time cycle referred to as the update cycle. It is typically 10 s.

Each switch i communicates the following traffic measurements to the network processor:[1]

I_{ij}—The number of idle trunks on the link to switch j, for all switches j in the DTM network.

CPU_i—The CPU occupancy of switch i.

O_{ij}—A measure of the traffic sent by i to j and which overflowed the direct route. j may be a switch inside the DTM network or some other resource that i can monitor, for instance a link to a switch outside the DTM network or to a customer.

These measurements are for both routing and congestion control purposes. For routing, they allow the network processor to build a global view of the idle capacity in the network. For congestion control, they allow it to protect potentially

[1]For reference purposes, a glossary of notations is provided in the Appendix.

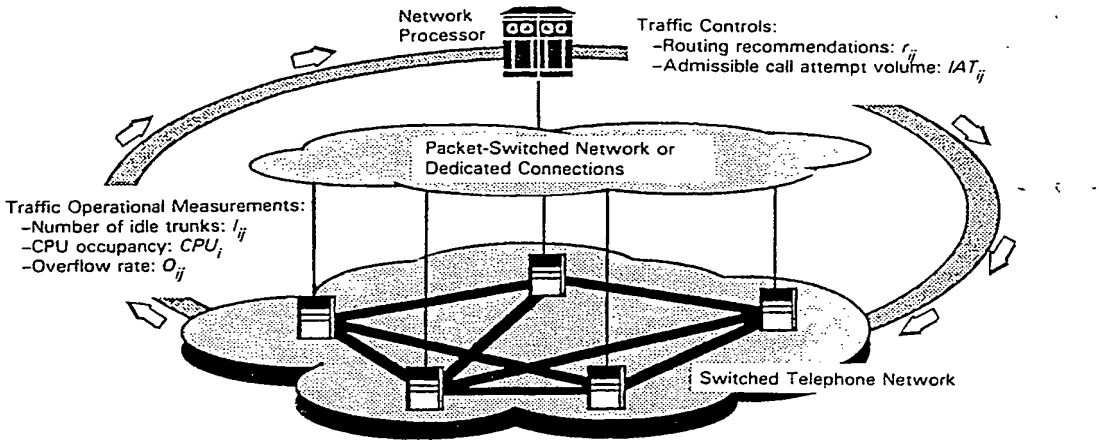


Fig. 1. DTM architecture and data flow.

overloaded elements in its control decisions.

Corresponding to the routing and congestion control functions, the network processor returns two types of controls to each switch i:

r_{ij}—For each switch j, the name of a switch through which i should redirect its calls to j when the direct route is full. r_{ij} may also be a "block" recommendation, in which case switch i blocks the calls to j that cannot be routed over the direct route.

IAT_{ij}—For each resource j, a minimum time interval between calls admitted to j. j may be a switch inside the DTM network or another resource that i can identify as a specific destination, for instance a link to a switch outside the DTM network or to a customer.

To simplify the discussion, it may be assumed that the above measurements and controls are systematically communicated every 10 s. The measurements, however, may in practice be filtered by the switches and only reported when needed. Similarly, the controls need not be systematically updated every 10 s, but only when the current controls are judged by the network processor to require and update. We will not dwell further on these implementation issues.

Routing

Consider a link of size N between two switches i and j, offered T Erlangs of Poisson traffic. Assume that each call which cannot be carried on the link is blocked, and that each blocked call costs one unit. Given that X trunks are currently busy, it has been shown [2] [3] that the expected cost for accepting an additional call on the link is:

$$c_{ij} = E[N,T] / E[X,T] \qquad (1)$$

where $E[X,T]$ is the Erlang-B formula for T Erlangs offered to X trunks. The cost in Equation 1 may be interpreted as the probability that the trunk held by the additional call during its lifetime would otherwise be used by a subsequent call, and causes this subsequent call to be blocked.

In a general telephone network, the expression for c_{ij} in Equation 1 may be modified to account for the impact of routing. In [4], the modification consists of multiplying the right-hand side of Equation 1 by a factor reflecting the alternate-routed traffic on the link and elsewhere in the network. This factor being typically close to unity, we shall for simplicity ignore it in the discussion. Assuming that the cost of an additional call on a multi-link route is the sum of its cost on each link, a near-optimal routing strategy for a general telephone network can be defined as follows [2] [3] [5]:

- Route an i-j call over:
  - —The direct route of $c_{ij} < 1$ and $c_{ij} < c_{it} + c_{tj}$ for all $t \neq i,j$
  - —The alternate route i-r-j if $c_{ir} + c_{rj} < 1$, $c_{ir} + c_{rj} < c_{ij}$ and $c_{ir} + c_{rj} < c_{it} + c_{tj}$ for all $t \neq i, j, r$
- Block the call otherwise (i.e., if $c_{ij} = 1$ and $c_{it} + c_{tj} \geq 1$ for all $t \neq i, j$)

This strategy assigns each call to the route with the minimal overall cost, provided that the overall cost is less than 1. The routes are limited to those with only one or two links. Longer routes, such as those with three or more links, are inefficient. In practice, they are avoided or used on an exception basis. The strategy also recommends blocking calls that cannot be carried over a route whose overall cost is less than 1. The rationale is that these calls cost more than they bring, i.e., one unit, and hence are not worthwhile.

The routing strategy in DTM is an approximation to the near-optimal strategy described above. The differences are motivated by implementation and fairness considerations. They aim at reducing the computation and communication burden for routing decisions, protecting against inaccurate network state and traffic knowledge, and reducing the discrepancies in the blocking of individual traffic parcels. We next describe the DTM routing strategy. Following this, we discuss its relationship with the near-optimal routing strategy.

Consider a call at switch i destined for switch j. In DTM, switch i always first attempts to route the call on its direct link to j. If the call is admitted into the network at switch i and if the direct link is full, switch i then attempts to route the call over the two-link alternate route transiting through the switch r_{ij}, as recommended by the network processor. This is the only alternate route that the call is allowed to attempt. If r_{ij} is the block recommendation or if the i-r link is full, switch i does not let the call attempt some other route but immediately blocks it. The alternate route recommendation r_{ij} applies only to calls originating at switch i, transit calls, i.e., calls arriving at i from some other switch for which i was the recommended DTM transit switch, may only attempt the direction route to j. Otherwise, these calls could take three or more links to complete, which would be inefficient.

The selection process of r_{ij} in the network processor depends on whether or not a direct link exists between i and j. If a direct link exists, which is the case for the vast majority of the calls, r_{ij} is determined as that switch t which achieves the maximum in:

$$\underset{t \neq i,j}{Max} \{ A_t \min [I_{it} - PA_{it}, I_{tj} - PA_{tj}] \} \qquad (2)$$

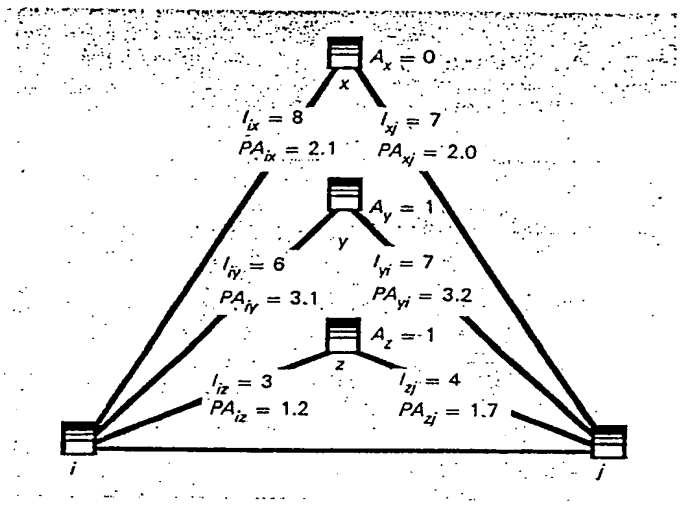


Fig. 2. Alternate route selection in DTM. Following Equation 2, the recommended alternate route is i-y-j. *Although it has the largest idle capacity, route* i-x-j *is not recommended because switch* x *is overloaded.*

provided that the maximum is positive. Otherwise, r_{ij} is the block recommendation. I_{ij} is as defined earlier. PA_{ij} is a protective allowance for the direct traffic on link *i-j*. Its role is to divert traffic away from the link when it is nearly fully occupied. We will return to the role of PA_{ij} later. A_t is a parameter in [0,1] reflecting the availability of switch *t*. It is 1 if switch *t* functions normally, but it is less if switch *t* is overloaded. Its role is to make alternate routes transiting through overloaded switches less attractive and hence less likely to be chosen by the network processor. We will also return to the role of A_t later, in the discussion on congestion control. Figure 2 illustrates the selection process of r_{ij} defined by Equation 2.

If *i-j* does not have a direct link, r_{ij} is determined as that switch *t* which achieves the maximum in:

$$\underset{t \neq i,j}{Max} \quad \{ A_t \ min \ [I_{it} , I_{tj}] \} \tag{3}$$

provided that the maximum is positive. Otherwise, r_{ij} is the block recommendation. Equation 3 is similar to Equation 2, except that the protective allowances are not considered. This is based on the rationale that the *i-j* traffic should not concede priority to the direct traffic on the links of its potential alternate routes when it has no direct link of its own.

Now consider the differences between the near-optimal and DTM routing strategies. It may first be noted that the near-optimal routing strategy implicitly assumes that routing decisions are made on a call-by-call basis and with exact knowledge of the idle capacities on all links. Such requirements are in practice very stringent. DTM instead makes routing decisions for groups of calls (namely, on a 10 s basis) and relies only on near-real-time knowledge of the idle capacities. The degradation entailed by this approximation to true real-time tracking and control motivates the first difference between the near-optimal and DTM routing strategies. Namely, DTM compensates by allowing calls to attempt two routes, the direct route and a recommended alternate route, rather than only one.

The systematic attempt of the direct route first may be explained by considering the cost c_{ij} in Equation 1 and the call set-up capabilities of the switches. Figure 3 depicts c_{ij} as a function of the number of busy trunks in a typical instance where $N = 100$ and $T = 80$ Erlangs. Clearly, c_{ij} is never greater than 1 and is close to 0 unless the number of busy trunks approaches N. This means that, irrespective of the number of calls in progress, it is never more "expensive" to let a call attempt its direct route than to block it. Considering that the direct route may be attempted by the originating switch without it losing control of the call, this makes it attractive as a first choice because the originating switch may still try another route when the direct route fails. In opposition, first attempting a two-link route precludes attempting the direct route upon congestion of the second link unless calls can be cranked back to their origin. Rather than imposing such a requirement, which may be difficult to justify economically or even to implement on several switching technologies, it is simpler to just let the calls first attempt their direct route. From the switching viewpoint, attempting the direct route first also reduces to the minimum the work of routing for the vast majority of calls. This is a nonnegligible side benefit considering that, with call translation, routing is a major source of switch real-time consumption.[2]

A second observation that can be made about the near-optimal routing strategy concerns the evaluation of c_{ij}. It is implicitly assumed that the traffic offered to the lines is known. This is not a trivial requirement. Traffic typically varies depending on the hour of day, the day of the week, the season, and on the routing itself. Knowledge of the traffic implies either continuous measurement or some kind of forecasting mechanism relying on historical data. Furthermore, even if the traffic is known, the evaluation of c_{ij} still involves significant computation (or memory if it is precomputed).

The above difficulties may be overcome by assessing the link cost with an approximation to c_{ij}. Figure 3 presents such an approximation, denoted c'_{ij}. It supposes that the cost of an additional call is either 0 or 1, depending on whether or not the number of busy trunks is less than or at least equal to a given threshold. The threshold, denoted t'_{ij}, summarizes the traffic information required for cost evaluation. This alleviates the difficulties for traffic estimation because t'_{ij} is relatively insensitive to traffic variations. c'_{ij} also simplifies the computations for assessing the link cost. It is merely a questions of comparing the number of busy trunks to the threshold.

Consider now the route evaluation in the near-optimal routing strategy with the approximate cost c'_{ij}. All routes for which the number of calls in progress on the first or the second link is at least equal to the link's threshold have a cost of at least 1, and hence are not recommendable. On the other hand, all other two-link routes have a cost of 0, and hence are all equally recommendable. Considering that measurements may potentially be only nearly accurate and that routing controls may have to apply to several calls, an auxiliary criterion that can help choose among the routes with 0 cost is robustness. From this viewpoint, and considering also that a two-link route must have both links available to be useful, the recommended route can be chosen as that with 0 cost and whose weakest link is as strong as possible. Namely, the recommended route would be that route *i-t-j* for which $c'_{it} + c'_{tj} = 0$ and $min\,[t'_{it} - X_{it}, t'_{tj} - X_{tj}]$ is maximum. Notwithstanding the factor A_t, whose purpose is congestion control, this selection process is precisely what Equation 2 implements. This can be seen by noting that $I_{ij} = N_{ij} - X_{ij}$ and by setting $PA_{ij} = N_{ij} - t'_{ij}$ and $A_t = 1$. This evidences a strong similarity between the DTM and near-optimal routing strategies. Namely, the DTM alternate route selection process may be viewed as an approximation to the near-optimal routing strategy, where the approximation serves to reduce the computational burden in the route selection and the impact of inaccurate knowledge of network and traffic conditions.

The protective allowance PA_{ij} is a means of diverting overflow traffic from the *i-j* link when it is very busy. This reduces

[2]Justification for attempting the direct route first may also be found in [6] and [7]. In addition, it was drawn to the authors' attention by Prof. Z. Dziong in a private conversation that attempting the direct route first may improve the near-optimal routing strategy. We hope that these results will be available shortly in the published literature.

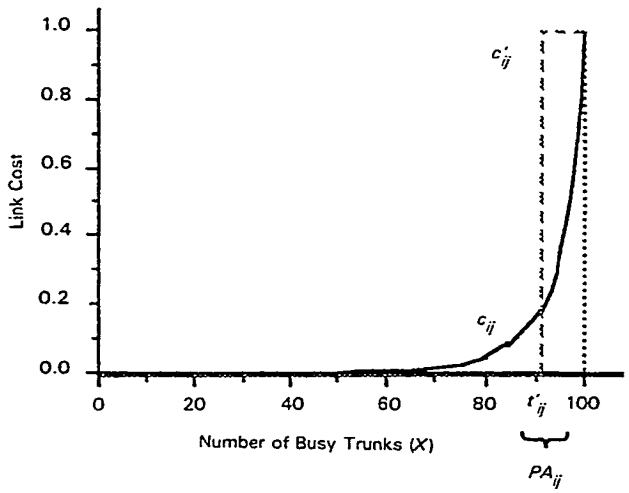


Fig. 3. True and approximate link cost functions.

the possibility that overflow traffic causes an inefficient usage of the trunk resources by taking away trunks that are likely to be needed for direct traffic. The protective allowance is a widely known mechanism used in several routing strategies [8–11]. Contrary to most routing strategies, however, the protective allowance in DTM is purely an algorithmic feature for biasing the routing decisions. It does not physically reside in the switches, which greatly simplifies implementation. To fully adapt to the current traffic load, the protective allowance in DTM may be dynamically computed based on near-real time measurements. This may be achieved, for instance, by using the overflow measurement O_{ij} [12]. Alternatively, in networks well-engineered for DTM, the protective allowance may be set to a fixed value with little loss in performance. We will return in more depth to this question later.

As a final observation, it may be noted that the criterion underlying the near-optimal routing strategy is the minimization of the overall blocking. This does not guarantee that all traffic parcels experience a satisfactory grade of service. A frequent situation in a typical well-engineered telephone network where this criterion may lead to poor performance is for traffic parcels that do not have a direct route. DTM makes the route computation slightly differently for these traffic parcels to provide them with a better chance of completion. Namely, DTM makes their route recommendation based on Equation 3 rather than Equation 2. This preferential treatment significantly reduces their blocking, although it may generally still remain higher than average. Furthermore, as traffic parcels with no direct route typically constitute a small fraction of the total traffic, their preferential treatment has a negligible impact on the overall blocking.

Congestion Control

Figure 4 depicts the typical behavior of a resource as a function of the arrival rate of work it has to perform. Initially, the occupancy of the resource increases linearly with the arrival rate. At a certain point, the resource becomes fully occupied and the occupancy saturates. Then, as the resource cannot accept additional work, the overflow rate from the resource starts increasing linearly with the arrival rate. In the region where the resource is saturated, the excess load that cannot be properly handled still imposes a burden, both on the resource itself and on the other resources involved in carrying it. This burden may seriously cut the net throughput if it consumes a significant amount of the resource's capacity. Preventing this is the purpose of congestion control.

The definition of "excess load" depends on the nature of the resource. For a switch, an arrival rate greater than the capacity cannot be accepted. This causes unacceptable dial tone and post-dial delays, and causes rapid degradation of the performance of other switches as well. For efficient operation, a switch must be maintained within the linear part of the occupancy curve, below the saturation threshold. In this respect, an appropriate metric for the excess load on a link can be obtained from the overflow curve. A congestion overflow threshold can be defined so as to maintain near-full utilization, and the excess load can be considered as any overflow that exceeds the threshold. This definition is illustrated in Figure 4.

For a link, the definition of excess load cannot be based on the same rationale as for a switch. A link is not adversely affected if it is offered an arbitrarily large number of call attempts. It simply blocks those calls which it cannot carry. The main motivations for restricting the call flow to a link is that the call flow may consume its bandwidth for signalling and may consume as well switch processing to get to the link. If the call flow is large and has a poor chance of completion, it may be worthwhile to restrict it at source to protect switches and to preserve bandwidth for effective attempts. In this respect, an appropriate metric for the excess load on a link can be obtained from the overflow curve. A congestion overflow threshold can be defined so as to maintain near-full utilization, and the excess load can be considered as any overflow that exceeds the threshold. This definition is illustrated in Figure 4.

To detect congestion, the network processor maintains an indicator of the load on each resource. This indicator, denoted L_r for resource r, is defined as follows:

$$\text{If } r = \text{switch } i\text{: } L_r = (1-\alpha_r)\, CPU_i + \alpha_r L_r^{old} \quad (4)$$

$$\text{If } r = \text{link } i\text{–}j\text{: } L_r = (1-\alpha_r)\, O_{ij} + \alpha_r L_r^{old} \quad (5)$$

Namely, L_r is a moving average of the occupancy or overflow rate of the resource, depending on its nature. The smoothing parameter α may in principle be tailored on an individual resource basis, but this is generally not necessary. Rather, it is sufficient to use a common α for all switches and another common a for all links. Note also that links in Equation 5 need not be only between switches in the DTM network. They can as well be to switches outside the DTM network or to large customers whose traffic unpredictability may warrant congestion control; for instance, ticket agencies or television stations.

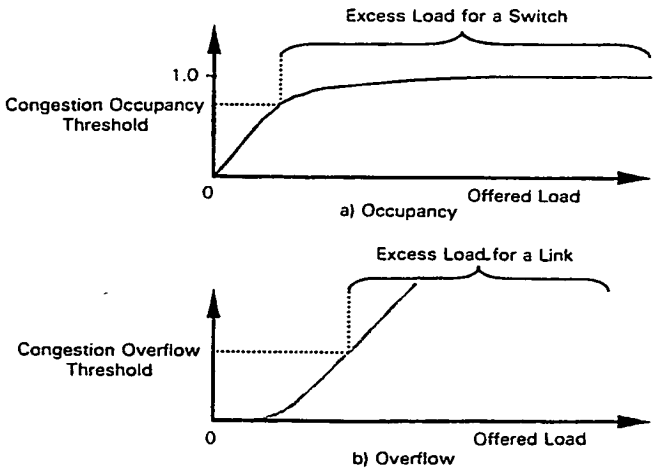


Fig. 4. Occupancy and overflow of a resource as a function of the rate of arrival of work into it.

When L_r for resource r exceeds a predefined threshold, denoted Ca_r, the network processor declares the resource under congestion and activates the congestion control process. Similarly, when L_r decreases below a second threshold, denoted Cd_r, the network processor considers the congestion over and deactivates the congestion control process. The activation and deactivation thresholds reside in the network processor. They can in principle be individually set for each resource but, as for α, it is typically sufficient to distinguish only on the basis of the nature of the resource. Also, activation thresholds must be set higher than deactivation thresholds so as to embed hysteresis in the process.

For switches, the network processor maintains also a third threshold, denoted Ch_r, for detecting when they are highly active. Ch_r must be lower than Ca_r but may for simplicity be set equal to Cd_r. The purpose of Ch_r is to allow gradual elimination of alternate-routed traffic from switch r as it approaches congestion. This is accomplished by means of the parameter A_r in the computation of the alternate route recommendations (see Equations 2 and 3). Namely, A_r is defined as follows:

$$A_r = \begin{cases} 1 \\ 0 \\ (Ca_r - L_r) / (Ca_r - Ch_r) \end{cases} \quad \begin{array}{l} \text{if } L_r < Ch_r \\ \text{if } Ch_r \le L_r < Ca_r \\ \text{if } Ca_r \le L_r \end{array} \tag{6}$$

As A_r multiplies the result of the minimization inside Equations 2 and 3, this makes switch r gradually less attractive as an alternate-route recommendation as L_r increases from Ch_r to Ca_r. When $L_r \ge Ca_r$, the switch is considered overloaded and is then fully freed from alternate-routed traffic.

Once the network processor declares a resource under congestion, it dispatches a control to every switch to limit the number of attempts they can let proceed into the network to the resource. This control, denoted IAT_{ij} for switch i to resource j, consists of a minimum interarrival time. It is implemented in switch i by accepting a call to j only if a timer exceeds the value of IAT_{ij}, and by resetting the timer to 0 every time a call to j is accepted. In its simplest form, new IAT_{ij} controls may be implemented in switch i only by updating the switch's current value for the control. Alternatively, more sophisticated implementations allowing better redistribution of the admitted call flow can also include the rescaling of the timer to immediately reflect the changes in the control. In all cases, the interarrival time control is "absolute" in the sense that the magnitude of the load accepted into the network is only weakly related to that of the offered load arriving from outside the network.

The computation of the interarrival time control relies on two key assumptions. The first regards the arrival rate from outside the network into the target resource. It is assumed that it is large, variable, and unpredictable. Hence, the control does not attempt to estimate it but rather to isolate the network from its wanderings. The second assumption regards the holding time of successful attempts to the target resource. It is assumed that this holding time is unknown, and that it is potentially different from the holding time under normal circumstances, but exists. This is motivated by the rationale that people have a well-defined purpose when they call in a situation of congestion. They may call to book tickets for a popular event, respond to a television contest, or enquire about relatives in the advent of a disaster; but a meaningful holding time can be associated with their behavior. Hence, the control can attempt to estimate the load per request on the congested resource, and tailor the overall rate of admitted calls to its capacity.

If the contested resource is r, IAT_{ir} for every switch i is determined as follows:

$$IAT_{ir}^{new} = S\, IAT_r^{new} \tag{7}$$

where:

$$IAT_r^{new}(instantaneous) = \begin{cases} CPU_j\, IAT_r^{old} / Ca_r, & \text{if } r = \text{switch } j \\ 10\ IAT_r^{old} / [\, IAT_r^{old}\ (Ca_r - O_{jk}) + 10\ ], & \text{if } r = \text{link } j-k \end{cases} \tag{8, 9}$$

$$IAT_r^{new} = Max[\,(1-\beta_r)\, IAT_r^{old} + \beta_r\, IAT_r^{new}(instantaneous),\ IAT_r(default)\,] \tag{10}$$

The superscripts *old* and *new* refer to the value of the variables before and after the update, respectively. S is the number of switches whose call attempt rate must be throttled to alleviate congestion. It is determined by the network processor by counting the number of switches whose overflow rate to r is significant, as reflected in the overflow measurements. $IAT_r(instantaneous)$ and IAT_r are immediate variables maintained by the network processor to estimate respectively the overall instantaneous and average interarrival time at r. B_r is a smoothing parameter for the averaging of IAT_r. $IAT_r(default)$ is a default for the overall interarrival time at r. It is set by assuming an optimistic (i.e., short) yet realistic holding time for successful call attempts at r. To prime the process, $IAT_r(default)$ is assigned as the first value to IAT_r.

Equation 8 assumes that as a switch (j in the equation), r behaves deterministically and that the overall arrival time IAT_r^{old} caused its CPU occupancy to be CPU_j, as measured. Considering a linear behavior for a switch as in the initial portion of the occupancy curve in Figure 4a, it determines a new value for the overall arrival time, denoted $IAT_r^{new}(instantaneous)$, by rescaling IAT_r^{old} to achieve the target occupancy Ca_r. Equation 9 similarly assumes that as a link (j-k in the equation), r behaves deterministically and that the interarrival time IAT_r^{old} caused its overflow rate to be O_{jk}. Considering a linear behavior for a link as in the high offered traffic portion of the overflow curve in Figure 4b, it determines a new value for the overall interarrival time to achieve the desired overflow rate Ca_r. Equations 8 and 9 are identical in spirit. The differences in their form result only from the different models used to predict the impact of interarrival time controls for switches and links.

Equations 8 and 9 both assume that the holding time per successful attempt is constant. This is motivated by the "meaningful holding time" assumption discussed earlier. They also assume that the target for congestion control is the congestion control activation threshold. Another target can in principle be defined, but there is little to be gained by adding this extra parameter. Equations 8 and 9 determine a new control based on only one measurement. Due to random fluctuations in the admitted call flow and holding time processes, this control may not always truly reflect the desired overall interarrival time. Equation 10 implements a moving average to filter out these random fluctuations and to prevent the overall interarrival time from drifting too far away from what can be recommended. It produces an estimate of the overall interarrival time, denoted IAT_r^{new}, which can be considered significantly more reli-

2460 IEEE Communications Magazine
28(1990)October, No.10, New York, US

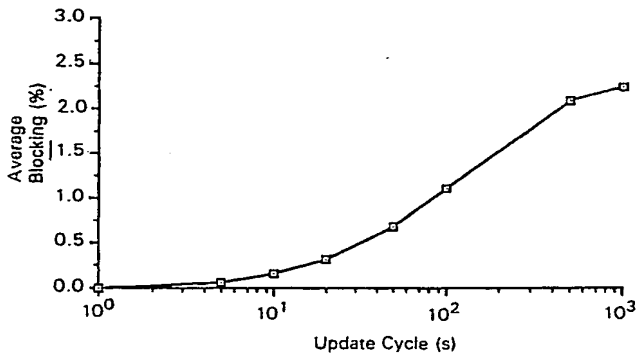


Fig. 5. Blocking as a function of the update cycle in a DTM network. The network contains 47 switches, each switch being on average directly connected to 21 switches, and 59,424 two-way trunks. Eight different nominal loads corresponding to several times of day and seasons are offered to the network. Their total traffics range from 35 to 45 KErlangs. The trunking is engineering for these eight nominal loads, assuming DTM. Each point on the figure is generated by simulating the operation of the network for 2.5 hours. The first 30 minutes are for initialization. The following 2 hours simulate sequentially each nominal load for 15 minutes. Exogenous traffics are modeled as independent Poisson processes, and call-holding times are modeled as independent, exponentially distributed random variables, with an average of three minutes.

able. Equation 7 straightforwardly apportions this overall interarrival time equally among all switches whose attempt rate warrants congestion control. Although the apportionment rule is very crude, this is of little concern because the feedback process can quickly correct the overall interarrival time to achieve the congestion control target.

Design Considerations

This section discusses the main motivations that have guided the design of DTM, as presented in the preceding section. As these motivations are covered to various extents in the existing literature, this discussion is not intended to be exhaustive. The interested reader may find abundant additional information in the references.

Update Cycle

Consider a large, realistic telephone network with, say, 40 nodes, 70% node-to-node connectivity, and handling 50 KErlangs. Well engineered under DTM, experience has shown that his network carries approximately 90% of the its traffic directly and has an overall trunk efficiency of approximately 75%. Hence, the network requires 50K/.75 = 67 Ktrunks, corresponding to an average link size of 61 trunks. Assuming an average holding time of 3 minutes, the network overflows 28 calls/s from their direct link, or 0.025 calls/s/link.

Now, aside from technical feasibility, the update cycle should be chosen so as to track, sufficiently quickly, the free capacity to allow overflow calls to make efficient use of the network. In view of the above network data and assuming that call overflow processes are perfectly regular, this consideration suggests that the update cycle could be set to 1/0.025 = 40 s. This would provide every overflow call with its own alternate-route recommendation. Considering, however, that overflow processes are typically not at all regular, that the above network data is only approximate, and above all that the network state may vary between the time the measurements are made and the time the resulting recommendations are applied, its is safer to set the update cycle significantly shorter, say, approximately 10 s.

As explained above, an update cycle of 10 s basically provides each overflow call with its own alternate-route recommendation. This eliminates the concentration of calls on individual alternate routes, which is one of the main causes of blocking. The other main cause of blocking is the evolution of the network state between the time at which the measurements are made and the time at which the resulting recommendations are applied. As discussed earlier, the approximate link cost function and the minimization in Equations 2 and 3 is a first means of countering it. Predicting the network state when the recommendations are enforced, and adjusting them accordingly, can be a second means of helping to achieve the same purpose. We have found that such predictions are difficult to make accurately, and that the additional time that they impose on the update cycle may defeat their whole purpose anyway. Finally, reducing the update cycle is obviously a third means of achieving better state knowledge. Here, we have found that the improvement is marginal unless the update cycle becomes unrealistically small.

Figure 5 depicts the evolution of the blocking as a function of the update cycle in a realistic DTM network operating under normal condition. Clearly, the blocking deteriorates seriously when the update cycle is significantly larger than 10 s, but it is relatively constant in the 5–10 s range. This confirms the incentive for lowering the update cycle to at least the 10 s range. For reasons of cost and technical feasibility, this confirms as well the incentive for not lowering it well below 10 s.

Protective Allowance

The cost function c_{ij} is a function of the traffic offered to the link *i-j*. For x given, $c_{ij}(x)$ increases as the traffic on link *i-j* increases. This can be reflected in DTM by making the protective allowance a function of the traffic as well. This has been investigated in [12]. There, the protective allowance is dynamically adjusted as a moving average based on the current traffic conditions via the formula:

$$PA_{ij}^{(n+k)} = \omega PA_{ij}^{(n)} + (1-\omega) M_{ij}^{(n,n+k)} \qquad (11)$$

where $PA_{ij}^{(n)}$ is the protective allowance at the nth update cycle, k is the interval in number of update cycles for the PA update computation, ω is a smoothing parameter for the averaging, and $M_{ij}^{(n,n+k)}$ is a measure of the number of trunks to be reserved for direct traffic computed from the overflow measurements received during the update cycles $n+1,\ldots n+k$. Equation 11 makes the link accessible to alternate-routed calls when the traffic on the link is small, but closes it when the link generates significant overflow. This is what the true cost function c_{ij} would also do.

To test the above dynamic update mechanism for the protective allowance, we made two series of simulations with modifications to the mechanism. In the first series of simulations, the protective allowance was forced to remain below an upper limit, denoted PA_{ij}^{max}, as follows:

$$PA_{ij}^{(n+k)} = Min\left[\omega PA_{ij}^{(n)} + (1-\omega) M_{ij}^{(n,n+k)}, PA_{ij}^{max}\right] \qquad (12)$$

In the second series of simulations, the protective allowance was forced to remain above a lower limit, denoted PA_{ij}^{min}, as follows:

$$PA_{ij}^{(n+k)} = Max\left[\omega PA_{ij}^{(n)} + (1-\omega) M_{ij}^{(n,n+k)}, PA_{ij}^{min}\right] \qquad (13)$$

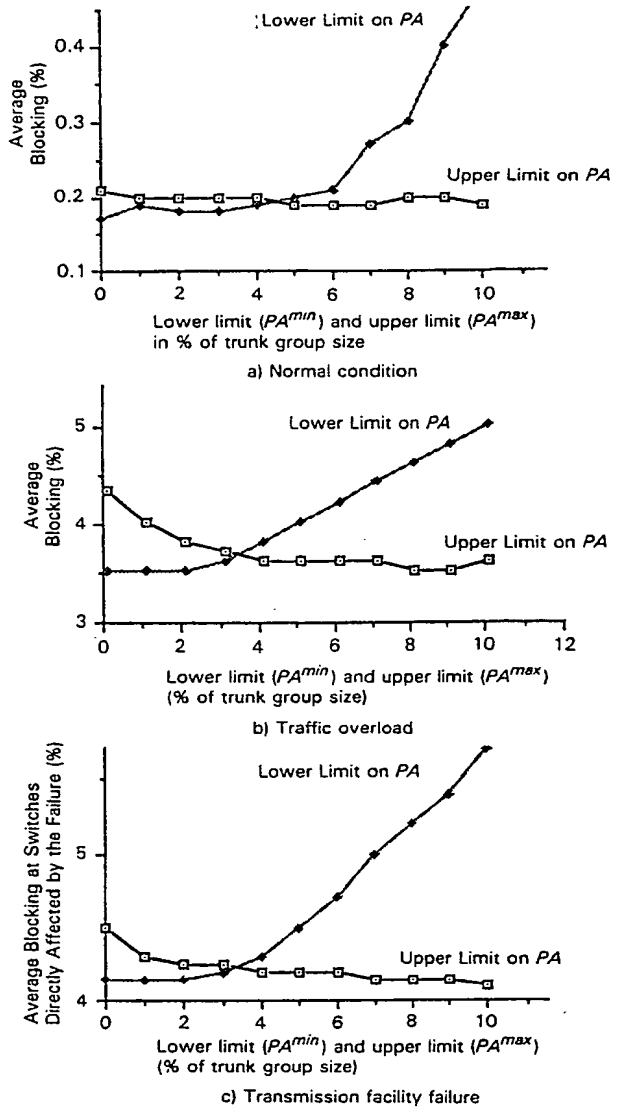


Fig. 6. Lower and upper limits on the protective allowance under normal condition, traffic overload, and transmission facility failure. The network and simulation conditions are identical to those of Figure 5 except for the following: In traffic overload condition, independent random fluctuations are imposed on every traffic parcel, inflating them on average by 20%. In failure condition, a large transmission facility failure is modeled as the loss of 17 trunk groups in a region of the network for an overall loss of 4,488 trunks.

Figure 6a shows the performance of the modified protective allowance update mechanisms in a realistic DTM network operating in normal condition. Two curves are presented. They depict the evolution of the blocking as a function of PA_{ij}^{min} and PA_{ij}^{max}. Concerning PA_{ij}^{min}, it may be seen that it has practically no impact if it is no more than 6%, but otherwise the blocking increases steadily as it increases. Concerning PA_{ij}^{max}, the blocking slightly increases when it becomes very small, but it has basically no impact. Figure 6b and 6c present the same curves as Figure 6a, but with the network operating respectively under traffic overload and under a large transmission facility failure. Although the blocking is higher due to the stress conditions, similar conclusions as those of Figure 6a apply. PA_{ij}^{min} and PA_{ij}^{max} have negligible impact if they are respectively no more and no less than 3%.

Two conclusions may be drawn from these simulations. First, the performance of the unconstrained protective allowance update mechanism is always at least as good as that of the constrained versions. Second, setting the protective allowance to 3% of the link size may provide nearly equivalent performance as that of the dynamic unconstrained mechanism. As it is simpler to implement, this may make the fixed protective allowance an attractive alternative when resilience to extreme stresses is not mandatory. It is worth noting that this conclusion corroborates the known fact that for physical trunk reservation, a protective allowance of 3% is adequate [13].

Multiple Alternate Routes

DTM recommends one alternate route per origin-destination pair. This is sufficient in a small network or with a nearly zero update cycle because the recommended routes may seldom fill before they get updated. In a large network with a 10 s update cycle, however, it may be argued that a second alternate route may help the traffic to complete, and hence reduce blocking. Allowing a second alternate route, however, is only useful if the originating switch still controls the call when the

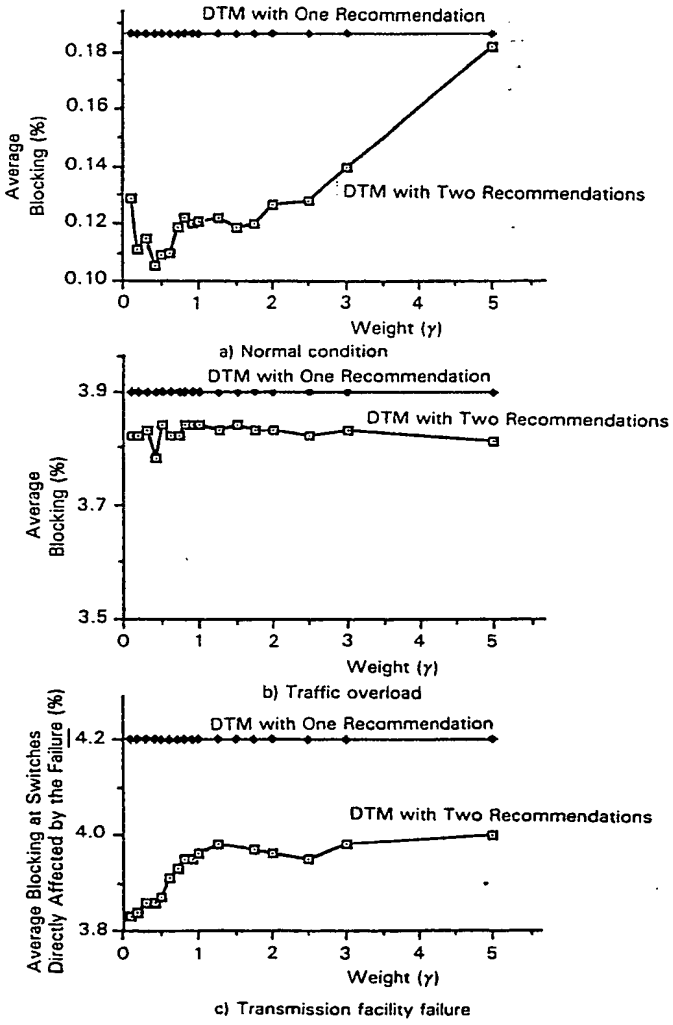


Fig. 7. DTM with two alternate-route recommendations per origin-destination pair in a network operating under normal condition, traffic overload, and transmission facility failure. The network and simulation conditions are identical to those of Figure 6.

first alternate route is busy. Otherwise, if the call is at the intermediate switch of the first alternate route, it is too late to attempt the second. Based on this observation, and in an attempt to gauge the benefits that an additional route could provide, we modified DTM so as to recommend two alternate routes. For *i-j*, the first alternate route *i-t-j* was chosen as that for which t achieves the maximum in:

$$Max \{ A_t \ min [I_{it} - PA_{it}, \gamma (I_{tj} - PA_{tj})] \} \quad t \neq i,j \tag{14}$$

$\gamma = 1$ corresponds to the standard DTM alternate route selection, as in Equation 2. $\gamma < 1$ makes the safe idle capacity on the second link more important than that of the first in the minimization inside the equation. This favors paths with more idle capacity on their second link than on their first, which are hence more resilient with respect to congestion on their second link. The second alternate route was chosen as in the standard DTM, but with the obvious restriction that it be different from the first.

DTM with two alternate-route recommendations was simulated in a realistic network operating in normal condition, under traffic overload and under a transmission facility failure. The results are presented in Figure 7. With $\gamma = 1$, the modification amounts to providing two standard recommendations instead of one. It is readily seen from Figures 7a and 7c that this lowers the blocking by 30% in normal condition and by 5% in failure condition. In addition, the blocking is further reduced by an additional 7% in normal condition and 8% in failure condition with a well-chosen γ. This is achieved when $\gamma \sim 0.4$. It is apparent that the enhancement is of little use in traffic overload. This is because the protective allowance mechanism then closes most links to overflow traffic. This results in calls seldom having access to two alternate routes, making the second recommendation useless.

Overall, the second alternate route significantly reduces the blocking only under normal condition. This is, however, where it is least needed. If it were worthwhile, such a gain could be achieved by adding slightly more trunks, say, 0.5%. Globally, this would be more cost effective as it would spare the network processor the significant overhead imposed by the computation of the second routes. Hence, as a global enhancement for all traffic parcels, it appears that the value of a second alternate route is limited.[3]

Similar results have been found in the study of three-link alternate routes. Namely, they offer a marginal advantage under normal condition and tend to degrade network performance under the slightest overload because they permit inefficient use of trunking resources. Moreover, considering them as potential routes requires a significant computational overhead compared to two-link alternatives: the time spent finding them is better spend by proceeding immediately to a new update cycle.

Congestion Control Thresholds

Let r be link *j-k*. Assume that under congestion, r maintains an efficiency of v_r and serves calls whose average holding time is H_r. If the overall interarrival time IAT_r^{old} was imposed on r, the overflow measurement O_{jk} that the network processor gets for the cycle should obey the equation:

$$O_{jk} = 10 / IAT_r^{old} - 10 v_r N_{jk} / H_r \tag{15}$$

Namely, the overflow from the link should be the difference between the number of calls it was offered ($10/IAT_r^{old}$) and the number of calls it was able to accept ($10 v_r N_{jk}/H_r$). This assumes that the update cycle is 10 s. The equation can be generalized easily to treat the update cycle as a parameter, but there little to be gained in adding this extra notation.

Based on Equation 15, the effective holding time of calls on the link, i.e., H_r/v_r, can be estimated as:

$$H_r / v_r = 10 N_{jk} / | 10 / IAT_r^{old} - O_{jk} | \tag{16}$$

Based on Equation 15, the new overall interarrival time control IAT_r^{new} can also be predicted so as to achieve the desired overall overflow target $O_{jk} = Ca_r$. Namely, IAT_r^{new} should satisfy:

$$CA_r = 10 / IAT_r^{new} - 10 N_{jk} / (H_r / v_r) \tag{17}$$

Assuming that H_r/v_r does not vary significantly from one update to the next, and using Equation 16 to evaluate it, this leads to defining IAT_r^{new} as:

$$IAT_r^{new} = 10 IAT_r^{old} / [IAT_r^{old} (Ca_r - O_{jk}) + 10 | \tag{18}$$

Note that the term inside the square brackets is always positive because $Ca_r > 0$ and because the overflow from the link may at most be equal to the number of calls offered to the link (i.e., $O_{jk} \leq 10/IAT_r^{old}$). Note also that Equation 18 requires that IAT_j^{old} be known, which is not the case when the equation is first used. This difficulty may be overcome by initially assuming an optimistic default value for H_r/v_r; for instance, $H_r/v_r = 10$ s. This default value can then be used to produce a starting default value $IAT_r(default)$ via Equation 17. It is important here that the default value be optimistic. Otherwise, the system could initially be too harsh, which could cause oscillations in the activation and deactivation of congestion control.

As it stands, Equation 18 has a major weakness. To illustrate it, suppose that the exogenous arrival rate to r is insufficient to achieve the desired overflow rate Ca_r. This may well happen under congestion because the exogenous arrival rate is unpredictable, hence may not be guaranteed to always be sufficient to achieve Ca_r. Equation 18 does not anticipate such a situation. When the overflow measurement is less than the desired target, it simply assumes that IAT_r is too harsh and, accordingly, reduces it to allow more calls to be accepted. This behavior, if it persists for several update cycles, may cause IAT_r to approach 0. If the exogenous arrival rate then suddenly increases, a significant number of update cycles may be required before Equation 18 can restore IAT_r to a proper value. Meanwhile, r will become vulnerable to high attempt rates.

The downward drift of IAT_r can be stopped by imposing a floor on it. For this purpose, it is natural as well as simple to reuse the assumption that the holding time on link r may possibly not be less than the default value used for initialization. This leads to defining IAT_r^{new} via:

$$IAT_r^{new} = Max\{ 10 IAT_r^{old} / [IAT_r^{old} (Ca_r - O_{jk}) + 10 | , IAT_r(default) \} \tag{19}$$

Note that this further strengthens the rationale for choosing H_r/v_r, hence $IAT_r(default)$, optimistically.

[3]This conclusion does not apply, however, to traffic parcels with no direct route. Then, the second alternate route may be an effective means to compensate for the lack of a direct route. We will not dwell further on this issue here, but it is worthwhile to remember the second alternate route as selective enhancement for traffic parcels with no direct route.

2460 IEEE Communications Magazine
28(1990)October, No.10, New York, US

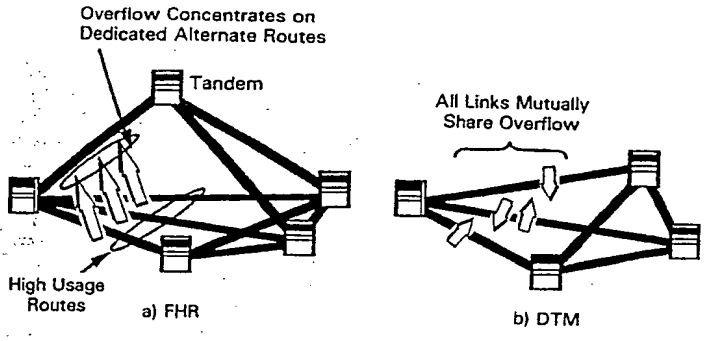


Fig. 8. Overflow paths.

The above observations constitute the rationale behind the definition of the overall interarrival time control for links in Equations 9 and 10. The rationale for switches is identical. The difference in the definition of the control in Equation 8 results only from the different nature of the measurement and congestion control target for switches—namely, occupancy as opposed to overflow—and in the specifics of the model relating them to the overall interarrival time control.

Algorithm Extensions

In many networks, there exist traffic parcels for which no direct link and no two-link route exist from the traffic's origin to destination switch. Also, the current mix of switching technologies may not provide all networks with the capability of implementing DTM at every switch. In both these cases, originating traffic parcels for which DTM cannot be applied may be routed, as today, according to a fixed overflow sequence until they encounter a DTM switch within two links of their destination, and then may be treated as locally originating by the DTM switch [14]. The routing rules described earlier ensure that these calls, upon reaching a DTM switch within two links of their destination, are completed over a remaining one- or two-link path, or else are blocked.

Rationale for Traffic Management Automation

This section outlines the immediate benefits that automation and near-real-time responsiveness entail in traffic management. These benefits are assessed through comparisons with today's routing and traffic management systems and processes. It is assumed that the reader is familiar with Fixed Hierarchical Routing (FHR) and the centralized manual traffic management systems typical of today's networks. Information on these may be found in [15] and [16] and, in general, in the proceedings of the International Teletraffic Congress and Networks conferences.

Capital Savings

Because they can be translated into hard economic figures, trunk savings have historically been the key consideration in the business cases developed for replacing FHR. These savings have unfortunately often been overemphasized to the detriment of other, less immediately quantifiable benefits. We discuss these savings in this section, but we hope with the following two sections to correct the perception that they alone constitute the rationale for automation of traffic management.

The trunk savings in DTM result from two main sources: the full sharing of trunk resources and the adaptation to predictable variations of the traffic load . Figure 8 explains the savings resulting from full resource sharing. Under FHR, high-usage routes may only carry their direct traffic. Considering the random nature of telephone traffic, these routes are typically provisioned to carry 85% of their direct traffic. Attempting to provide more trunks to directly carry more traffic is in general uneconomical because the additional trunks would seldom be used. Rather, it is preferable to let the traffic overflow onto a small number of tandems where, as a result of concentration, trunks can be efficiently used. Under DTM, high usage routes may all carry alternate-routed traffic. As a result of this full sharing, the routes can be provisioned to a higher number of trunks for a given efficiency. This, in turn, increases the proportion of traffic completing directly, and thereby reduces the proportion of traffic completing over two or more links. Overall, routes in DTM have a comparable efficiency to that of high-usage routes in FHR, but they typically let only 5–10% of the direct traffic overflow. Comparing this figure to the 15% for FHR and noting that overflow traffic requires at least two links, this results in a 5–10% trunk savings [17] [18].

An important yet often forgotten consequence of the above trunk savings is the associated savings in switch utilization. Namely, a direct call visits two switches while a two-link call visits three. By directly routing 5–10% more calls, DTM reduces by 3–5% the number of switch attempts that calls generate. In local networks where switching is roughly twice as expensive as trunking, this translates into additional savings a significant as the trunk savings themselves.

Figure 9 explains the savings that can be achieved by adapting the routing to predictable variations of the traffic load. The figure depicts a simple three-node network with two distinct busy hours. The first busy hour consists of traffic from node 1 to 2 and 2 to 3 only, and the second of traffic from node 1 to 3 only. Under DTM, the links from node 1 to 2 and 2 to 3 can carry the traffic from node 1 to 3 in the second busy hour. Under FHR, unless node 2 is the "home" of node 1 or 3, trunks must be provisioned between nodes 1 and 3 to support the traffic during the second busy hour. As in general very few nodes may serve has home, this inflexibility leads to over provisioning. In practice, the savings associated with adaptation to predictable variations of the traffic load depend on the noncoincidence of local traffic peaks. They may reach 2–4% in metropolitan networks with distinct residential and business busy hours, and 4–6% in toll networks spanning several time zones.

The above savings are not specific to DTM. They are basically achieved by all routing strategies allowing full resource sharing and adapting to predictable variations of the traffic load. However, compared to other routing strategies proposed to replace FHR, DTM does not require preplanned optimization of the routing based on historical data to achieve these savings. This greatly simplifies the off-line support for routing, and ensures efficient operation when operation conditions differ from those for which the network is engineered.

Traffic Management Automation

In FHR, the routing architecture is physically engineered in the switches, based on off-line optimization and historical

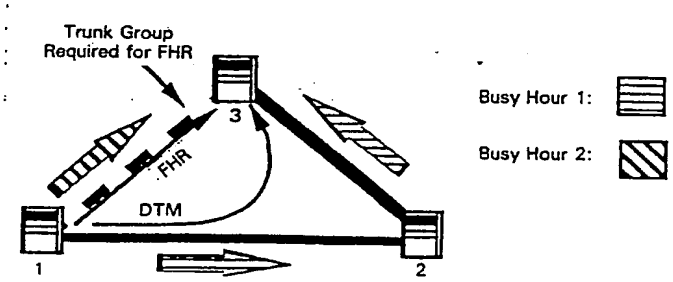


Fig. 9. Taking advantage of noncoincident busy hours.

2460 IEEE Communications Magazine
28(1990)October, No.10, New York, US

Table I. Manual Traffic Management Controls

Traffic Management Control	Relative Frequency of Activation (%)	Type of Action
Reroute	45	Expansive
Cancel From/To	28	Restrictive
Code Blocking	12	Restrictive
Traffic Overload Reroute Control (TORC)	9	Expansive
Skip Route	3	Expansive
Others	<3	Expansive and Restrictive

data. It has no possibility of adaptation to traffic patterns or load on the equipment. Networks, however, must withstand traffic- and equipment-related stresses for which they cannot *a priori* be provisioned. Some of these stresses, such as on Mother's Day and Christmas, are predictable; but their exceptional nature cannot warrant their inclusion in provisioning considerations. Other stresses, such as equipment failures or mass calling to an area struck by a natural disaster, just cannot be anticipated.

To cope with exceptional or unpredictable traffic stresses, most networks are supervised by (manual) centralized traffic management systems. These systems collect operational measurements on resource status and traffic patterns and allow remote activation of network elements traffic controls (see, for instance, [16]). The operational measurements are analyzed by the systems, typically via comparison with preset thresholds, and alarms are raised when abnormal conditions occur. Network managers correlate these alarms to their potential sources and enforce contingency plans when required. The controls used for traffic management may be either restrictive or expansive. Restrictive controls bar, at the source, traffic from entering the network. They are used to protect the network integrity when exceptional traffic or network element stresses occur. Expansive controls override or modify the routing plan implemented in the switches to permit redirection of traffic flows. They are used in conjunction with restrictive controls to redirect traffic away from congested areas, or when the traffic differs significantly from its forecast but can yet be served. Table I summarizes the main traffic controls used today together with their relative frequency of activation.

DTM, on the other hand, is a traffic management system. It automatically tailors the routing to the current switch loads and traffic patterns and restricts access to the network whenever required. This differs from current traffic management in three main aspects.

First, human intervention is eliminated from the active control process. Human intervention is required to set the behavior of the control process—for instance, to set thresholds for congestion control activation or to define admissible routing paths—but it is not part of the control process itself. This allows the control cycle to be on the order of seconds instead of several minutes, as is typical today.

Second, the short update cycle effectively institutes a real-time feedback loop between the network processor and the controlled network. This real-time feedback loop enables the traffic management control process to rely on simple measurements and control algorithms. Today, collection of extensive operational measurements and complex preplans are justified because control decisions must be carefully made. This is sound when decisions may remain enforced for a significant time before new data and human intervention can allow corrections. Extensive operational measurements and complex preplans, however, are not required when control decisions can be quickly corrected based on feedback.

Third, the switches dispatch to the network processor direct measurements of the end-to-end traffic. This provides the network processor with immediate knowledge of the traffic demand. Today, traffic measurements are made on a trunk group basis. When traffic pressure occurs, network managers must reconstruct the end-to-end traffic demands, creating them from trunk group measurements, before they can take corrective actions. Directly providing end-to-end traffic measurements avoids this reconstruction process. This simplifies the control process in the network processor, thereby allowing it to be fast, and eliminates the requirement for human correlation of link to end-to-end traffic.

Figures 10 and 11 illustrate the traffic management capabilities of DTM. Figure 10 depicts the evolution of the blocking, trunk utilization, and alternate-route selection process in a DTM network subjected to a large and sudden transmission facility failure. The blocking increases when the failure occurs because the affected switches suddenly do not have enough capacity to let their traffic out. The blocking, however, quickly stabilizes. Then the remaining working trunks at the switches affected by the failure are utilized very efficiently (see Figure 10b). The bulk of the traffic that cannot complete is barred at the source from the network via the "block" recommendation, and the traffic at the switches not directly affected by the failure is diverted from the congested area (see Figure 10c). In other words, DTM, in less than a minute, implements all the control actions necessary to maintain maximum efficiency at

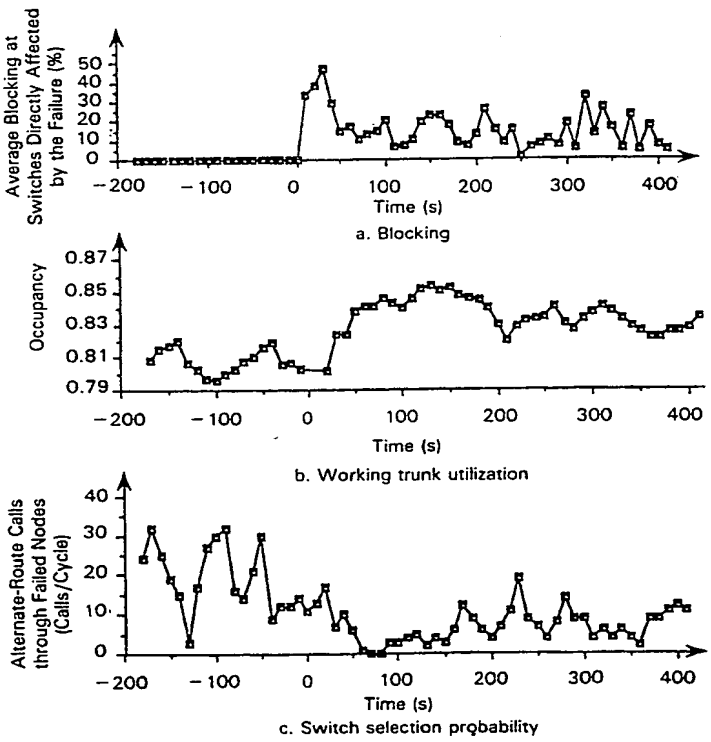


Fig. 10. Transmission facility failure in a DTM network. The failure occurs at time t = 0. The figure depicts the blocking at the switches directly affected by the failure, the utilization of the working trunks at the switches directly affected by the failure, and the probability of selection of the switches directly affected by the failure as alternate routes by the other traffic parcels. The network is identical to that of Figure 5 except for the following: The transmission facility failure is modeled as the loss of 80% of the capacity of 4 large links for an overall loss of 1,854 trunks. The failure occurs after 30 minutes of initialization and under the highest traffic load.

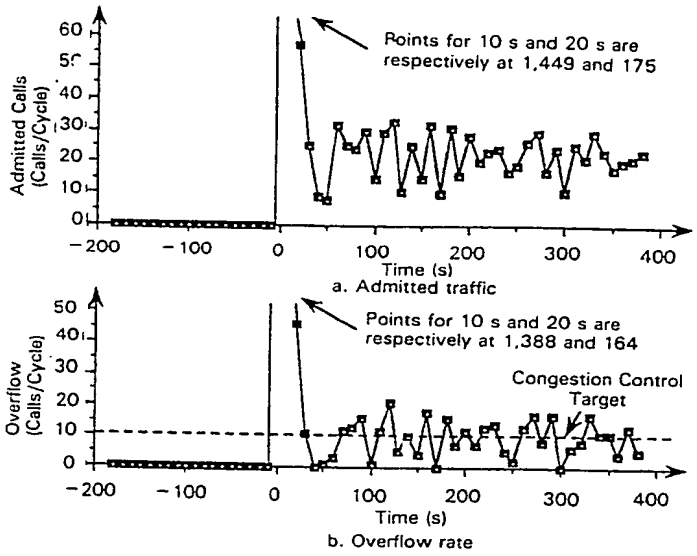


Fig. 11. Mass focused calling to a target customer in a DTM network. At time t = 0, the traffic to the target customer increases from 0 to 5 KErlangs. The figure depicts the traffic admitted into the network to the target customer and the overflow rate from the link to the target customer. The link congestion activation threshold is 10 calls/update cycle. The holding times of successful attempts to the target customer are modeled as independent, exponentially distributed random variables with an average of 36 s. The holding time assumed as the default is 10 s. The network is identical to that of Figure 5. The mass calling occurs after 30 minutes of operations under normal condition, and under the highest traffic load.

the switches affected by the failure and to protect the other traffic in the network.

Figure 11 depicts the evolution of the traffic load admitted to a specific link and the overflow from the link in a DTM network when the link is subjected to a sudden intense traffic stress. In this example, the link connects to a customer outside the DTM network and its offered traffic suddenly increases from 0 to 5 KErlangs. As the link cannot possibly sustain this traffic, the network must block, at the source, as much as possible of the excess traffic. This prevent it from assaulting the switch on which it homes or propagating to other switches in trying to complete, and thereby protects the other traffic in the network. Figure 11a shows that DTM, in less than a minute, throttles the traffic admitted to the link to the level required to achieve the desired overflow rate (see Figure 11b). Note also that DTM enforces precisely the congestion overflow rate target preset in the network processor.

Trunk Servicing

Trunk servicing is the support activity that ensures that the trunk network meets its performance objectives under normal day-to-day conditions. It consists of monitoring the network and initiating short-term corrective trunk provisioning actions whenever the evolution of the traffic demand can lead to unacceptable degradation of performance. Before we assess the implication of DTM on servicing, we first present an experiment that will help motivate the discussion.

Let p_{ij} be a perturbation of the size of link i-j, initialized as follows:

$$p_{ij} = N_{ij}(1 + p) \quad \text{if } i+j \text{ is even}$$
$$p_{ij} = N_{ij}(1 - p) \quad \text{if } i+j \text{ is odd} \qquad (20)$$

p is a parameter controlling the amplitude of the perturbations. Now let the p_{ij} be adjusted as follows:

$$p'_{ij} \leftarrow p_{ij}(\text{Sum}_j\, n_{ij} / \text{Sum}_j\, p_{ij}) \quad \text{For all } i, j$$
$$p''_{ij} \leftarrow p'_{ij}(\text{Sum}_i\, n_{ij} / \text{Sum}_i\, p'_{ij}) \quad \text{For all } i, j \qquad (21)$$
$$p_{ij} \leftarrow \text{Integer}(p''_{ij}) + 1 \quad \text{For all } i, j$$

and let this adjustment be iteratively repeated until all row and column totals of the matrices $[N_{ij}]$ and $[p_{ij}]$ are within the limits of a stopping criterion (set to 0.5% in the example that follows). In practice, this procedure always converges for realistic and reasonable size matrices with, say, at least 50 non-zero entries and of size at least 10 × 10. In fact, the perturbations in Equation 20 typically cancel each other out, so the iterations in Equation 21 result only in minor adjustments. Overall, the procedure produces a matrix $[p_{ij}]$ with the same outgoing and incoming nodal trunking (i.e., row and column totals) as $[N_{ij}]$, but with individual trunking parcels distorted by factors of $+p$ and $-p$.

Consider now a DTM network nominally engineered for the link matrix $[N_{ij}]$, but suppose that the network is instead imposed the perturbed link matrix $[p_{ij}]$. This should worsen the grade of service because the network will not be adapted as well to its traffic. The severity of the degradation will depend on both the magnitude of the perturbations and the ability of the routing to cope with them. Figure 12 presents the result of such an experiment. It depicts the evolution of the blocking as a function of p in a typical DTM network. Clearly, the figure demonstrates that DTM can sustain perturbations of up to 20% before performance starts degrading.

The behavior of DTM in Figure 12, as well as a comparison with FHR, may be explained by considering Figure 8. In DTM, overflows from undersized links may be allocated onto all links. In this respect, they can benefit from the idle capacity on all oversized links, and hence may still be provided with a good chance of completion. However, as p increases, the traffic completing over two-link routes increases. This causes the network to become less efficient and eventually leads to a degradation of performance. Quantitatively, the degradation becomes noticeable when p is about 20%. By contrast overflows from undersized links in FHR would only have access to a very small number of tandems. Although oversized links would generate less overflow, the overall result would be additional traffic to the tandems. As the tandem links are not engineered for it, the net result would be a linear increase of the blocking as a function of p.

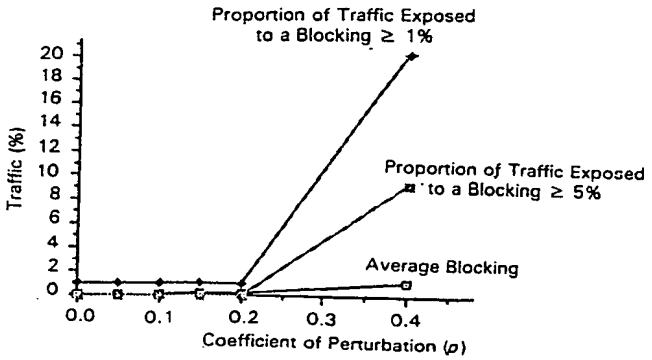


Fig. 12. Blocking in a DTM network as a function of the coefficient of perturbation p.

The key observation that can be made from the above experiment is that, given constant nodal trunking, large perturbations in link sizes have no impact on grade of service in a DTM network. This has three main implications for servicing [19].

The first implication is on the frequency of servicing. Typically, the overall traffic forecast for a switch is significantly more stable than that for its individual traffic parcels. For instance, if individual traffic parcels can be considered as independent identical random variables, the probability that the overall traffic forecast for a switch exceeds its expected value by a given factor is roughly $k^{0.5}$ smaller than that for its individual traffic parcels, where k is the number of traffic parcels on the switch. Under FHR, parcel deviations exceeding their forecast individually trigger servicing, as they would otherwise overload their direct link or the link to some tandem. Under DTM, these deviations require servicing only if the overall nodal traffic increases, which is $k^{0.5}$ less frequent.

The second implication is on the granularity of servicing. In FHR, trunks added on a direct link benefit only the traffic parcel using the link. In DTM, trunks added on any link benefit several traffic parcels. This allows servicing to be on a bulkier, hence less frequent, basis. Indeed, instead of adding or displacing small amounts of trunks as is typical today, servicing in DTM can be done at a larger modularity. Even if this sometimes means installing more trunks than are strictly required, the additional trunks are worthwhile. They can quickly be used to absorb traffic deviations elsewhere in the network, thereby eliminating the servicing actions that these deviations may otherwise require.

The third implication is on the overprovisioning requirement for maintaining an adequate grade of service. Typically, networks are overprovisioned to provide a safety margin during which servicing can provision resources before the grade of service becomes unacceptable. As resources are better shared, this overprovisioning need not be as high under DTM as under FHR. For instance, using a similar argument to that above, a given grade-of-service margin would require roughly $k^{0.5}$ less overprovisioning under DTM than under FHR.

Overall, DTM makes trunk servicing a switch-based rather than link-based function. This allows it to be bulkier, hence more effective, and an order of magnitude less frequent than today.

Conclusion

This article has presented and explained the key design decisions of the DTM system being deployed in Canadian telephone networks. Distinguished from other routing methods by its use of global network status information for routing and flow control updates, and by its very short update period, it merges the planning of traffic routing with traditional real-time network management into a single automated system. Its use of near-real-time feedback eliminates the need for assumptions or measurements of prior traffic distributions to determine optimal traffic routing, and eliminates the need for extensive operational measurements, complex preplans, and active human involvement for real-time traffic management in the event of traffic stresses and equipment failures. Lastly, its simplicity enables straightforward implementation as well as extension to networks of mixed switching vintages and modes.

Acknowledgments

The work presented in this article stems from ten years of research at Bell-Northern Research (BNR), during which many individuals have made significant contributions. Although it is impossible to mention them all here, we certainly do want to gratefully acknowledge their contributions. We must mention, however, Messrs. M. E. Lavigne and W. J. Graham from Bell Canada for their continued and indeflectible support, and for the insight which they have constantly demonstrated in guiding the evolution of DTM. We must also mention Mrs. F. Caron and Mr. F. Langlois from BNR for their recent contributions and their help with this article.

The authors also gratefully acknowledge the continued support of Bell Canada for the project. This is, in the end, what has made the vision first published in 1967 [1] now a reality.

Appendix: Glossary of Notations

The following is a list of the notations recurring throughout the article. For the sake of brevity, notation used only locally is omitted.

A:	Switch availability parameter	*IAT*:	Interarrival time control
c:	Link cost function	L:	Load indicator
c':	Approximation to c	N:	Link size
Ca:	Congestion activation threshold	O:	Overflow rate
Cd:	Congestion deactivation threshold	PA:	Protective Allowance
Ch:	Switch high activity threshold	r:	Alternate route recommendation
DTM:	Dynamic Traffic Management	S:	Number of switches
FHR:	Fixed Hierarchical Routing	T:	Traffic
H:	Holding Time	X:	Number of busy trunks
I:	Number of idle trunks		

References

[1] C. Grandjean, "Call Routing Strategies in Telecommunication Networks," *ITC 5*, New York, NY, 1967.

[2] T. J. Ott and K. R. Krishnan, "State-Dependent Routing of Telephone Traffic and the Use of Separable Routing Schemes," *ITC 11*, Kyoto, Japan, 1985.

[3] Z. Dziong, M. Pioro, U. Korner, and T. Wickberg, "On Adaptive Call Routing Strategies in Circuit-Switched Networks—Maximum Revenue Approach," *ITC 11*, Kyoto, Japan, 1985.

[4] K. R. Krishnan and T. J. Ott, "Forward-Looking Routing: A New State-Dependent Routing Scheme," *ITC 12*, Torino, Italy, 1988.

[5] V. G. Lazarev and S. M. Starobinets, "The Use of Dynamic Programming for Optimization of Control in Networks of Commutation of Channels," *Eng. Cyber.*, no. 3, USSR Academy of Sciences, 1977.

[6] P. Chemouil, J. Filipiak, and F. Gauthier, "Analysis and Control for Traffic Routing in Circuit-Switched Networks," *Comp. Networks and ISDN Syst.*, vol. II, 1986.

[7] J. Filipiak, *Modelling and Control of Dynamic Flows in Communication Networks*, Springer-Verlag, 1988.

[8] G. T. Ash, R. H. Cardwell, and R. P. Murray, "Design and Optimization of Networks with Dynamic Routing," *Bell Syst. Tech. J.*, vol. 60, 1981.

[9] J. M. Akinpelu, "The Overload Performance of Engineering Networks with Nonhierarchical and Hierarchical Routing," *ITC 10*, Montreal, Canada, 1982.

[10] R. J. Gibbens, F. P. Kelly, and P. B. Rey, "Dynamic Alternate Routing—Modeling and Behavior," *ITC 12*, Torino, Italy, 1988.

[11] A. Inoue, H. Yamamoto, and Y. Harada, "An Advanced Large-Scale Simulation System for Telecommunication Network with Dynamic Routing," *Networks*, Palma de Mallorca, Spain, 1989.

[12] W. H. Cameron, "Simulation of Dynamic Routing: Critical Path Selection Features for Service and Economy," *Int'l. Conf. on Commun.*, Denver, CO, 1981.

[13] G. T. Ash, "Use of a Trunk Status Map for Real-Time DNHR," *ITC 11*, Kyoto, Japan, 1985.

[14] W. H. Cameron, J. Regnier, P. Galloy, and A. M. Savoie, "Dynamic Routing for Intercity Telephone Networks," *ITC 10*, Montreal, Canada, 1982.

[15] J. C. Truitt, "Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks," *Bell Syst. Tech, J.*, vol. 33, pp. 421–514, 1954.

[16] "Network Traffic Management (NTM) Operations System (OS) Requirements," *Bellcore Tech. Adv.* TA-TSY-000753, issue 2, 1988.

[17] E. Szybicki and A. E. Bean, "Advanced Traffic Routing in Local Telephone Networks: Performance of Proposed Call Routing Algorithms," *ITC 9*, Torremolinos, Spain, 1979.

[18] W. H. Cameron, P. Galloy, and W. J. Graham, "Report on the Toronto Advanced Routing Concept Trial," *Telecommun. Networks Planning*, Paris, France, 1980.

[19] M. E. Lavigne and J. R. Barry, "Administrative Concepts in an Advanced Routing Network," *ITC 9*, Torremolinos, Spain, 1979.

Biography

Jean Regnier is manager of a research group at BNR focusing on the modernization of telecommunication network operations. His current research interests include traffic management and personal mobile communication services in telecommunication networks.

W. Hugh Cameron manages network operations architecture, strategy, and system design groups in BNR.

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 400 879 A2**

(12) EUROPEAN PATENT APPLICATION

(21) Application number: **90305585.3**

(51) Int. Cl.[5]: **H04Q 3/66, H04Q 3/00**

(22) Date of filing: **23.05.90**

(30) Priority: **30.05.89 US 359015**

(43) Date of publication of application:
05.12.90 Bulletin 90/49

(84) Designated Contracting States:
DE FR GB IT

(71) Applicant: **AMERICAN TELEPHONE AND TELEGRAPH COMPANY**
550 Madison Avenue
New York, NY 10022(US)

(72) Inventor: **Gordon, Travis Hill**
41 Winding Way
Madison, New Jersey 07940(US)

(74) Representative: **Buckley, Christopher Simon Thirsk et al**
AT&T (UK) LTD. AT&T Intellectual Property Division 5 Mornington Road
Woodford Green, Essex IG8 OTU(GB)

(54) Dynamic shared facility system for private networks.

(57) This invention relates to a dynamically shared facility network (DSFN) providing private network service to a plurality of customers using switched facilities of a common carrier network. A plurality of serving offices are connected via access links to customer telecommunications equipment. A pool of channels is dedicated to providing communications for private network service among these serving offices. In response to a request from a customer, connections are set up in the serving offices between access links and members of the pool of channels, in order to interconnect the serving links sought to be connected by the request. Where tandem connections between serving offices are necessary, connections are set up between members of the pool of channels. In case of failure of one or more channels, a new connection is automatically established. Advantageously, communication channels of the large communications facilities of a public switched network can be allocated to the DSFN, thus achieving economies of scale, and thus permitting use of the large and flexible switching systems of the public switched network to control and switch channels of the DSFN.

EP 0 400 879 A2

DYNAMIC SHARED FACILITY SYSTEM FOR PRIVATE NETWORKS

Technical Field

This invention relates to arrangements for providing telecommunication service to private customer networks.

Problem

In recent years there has been a rapidly increasing demand for large private data networks to connect, for example, a large number of terminals such as reservation terminals to central host computers used for keeping track of reservation data. In addition, there has also been an increase in the voice private network services required by companies which are geographically dispersed, to handle the increasing volume of voice traffic among branch locations and between branch and headquarters locations. Private voice networks exist primarily as economically attractive alternatives to public network use. Data networks are implemented on private facility networks because prevalent host computer applications assume continuous connectivity to terminals. This characteristic makes the current public telephone network, arranged to provide transmission for many short communications, unsuitable and excessively coslly for these applications.

Virtual networks, such as AT&T's Software Defined Network (SDN), set up call connections one at a time, in response to a dialed request from a caller, over the public switched network, while providing customer features, such as special in-network dialing arrangements. Such networks, whose callers compete for service with the general public, do not provide the very high availability of dedicated private networks needed, for example, for inter-computer data transfers, necessary for the orderly conduct of a business.

Increasingly, customer dedicated digital facilities such as those provided by the Accunet® T1.5 service offered by AT&T are used as the private network backbone facilities between major customer locations. A digital facility consists of an access link between the customer premises and an AT&T serving office and an interoffice link provisioned from digital carrier systems connecting AT&T offices. Multiplexers in the major customer locations are used to derive both voice and data circuits fiom the digital facility. Such an arrangement is illustrated in FIG. 1. Facilities are provisioned by a common carrier through static cross-connect arrangements such as the Digital Access and Cross Connect System manufactured by AT&T Technologies, in the serving office. For high bandwidth facilities such as those offered by Accunet T1.5 Service these cross-connections are manually patched between the carrier terminal equipment of digital carrier systems. In this manner a common facility is shared among applications. These networks have become more complex as more customers have become geographically diverse and as the use of point to point facilities has become less economical and is being replaced by the use of networks providing switching arrangements. Typically, private network locations are not fully interconnected by digital facilities. Then, two locations may be interconnected via one or more tandem locations, implemented through the use of private branch exchange (PBX) facilities and specialized flexible multiplexer systems, in order to use the private digital facilities efficiently. A switching or cross-connect junction is then required on the customer's premises. Efficient use of such arrangements requires individual circuits to traverse the least number of tandem points. As these networks become more congested such routing is not always possible and over time inefficient routing develops which requires periodic administrative rearrangement.

Increasingly, customers have come to depend on these communication facilities in order to carry out their routine work so that the reliability of these network facilities has become critical. The engineering of networks and the administration of networks to provide this high reliability is complex and expensive. Further, following failures in the network and the use of backup facilities in response to these failures, restoration of these networks to the normal traffic carrying pattern is a complex and time consuming task. Accordingly, a problem of prior art large private customer networks is that high costs for network engineering, operations and administration are incurred in providing highly reliable service in the presence of network failures and frequently changing traffic pattern demands of the customers.

Solution

The above problems are solved and an advance is made in the state of the art in accordance with the principles of this invention, wherein, illustratively, a telecommunications network comprising a plurality of switching systems and interswitch transmission facilities comprises a pool of interswitch communication channels dedicated for use by a plurality of private networks. Each private

network is connected to ones of a plurality of toll switches by access facilities. These access facilities define and limit the use of interswitch channels; the private customers are allowed to set up on demand any set of interconnections among their access facilities, the interconnections being provided by the pool of channels of the network. In response to a request data message from a customer administrator of one of the private networks, any interswitch channel(s) that is currently available may be assigned for connecting the access facilities specified in the message. A data base maintains a record of use of access facilities by the private customer to ensure that the private customer does not exceed his allotted capacity. Each switch maintains a record of the trunk groups and the busyidle status of all the channels (trunks) of the pool that are connected to that switch and maintains a routing data base for selecting an optimum route for connecting the access facilities to be connected in response to any request. The pool is engineered to provide sufficient transmission facilities to interconnect all the access facilities of the private network customers in any combination, and to provide an adequate number of extra facilities to be used in case of failure of one or more of the facilities in the pool. This permits each of the customer admimstrators to draw facilities from the pool without exhausting the pool and without requiring permission from a network administrator.

In accordance with one aspect of the invention, a shared public network also used for public telecommunications service is used to provide facilities for such a dynamically shared facility network (DSFN). Advantageously, communications channels of the large communications facilities of the public switched network can be allocated to the DSFN, thus achieving economies of scale. Advantageously, the large and flexible toll switching systems of the public switched network can be conilgured to control and switch the channels of the DSFN. Advantageously, such sharing eliminates the need for tandem points in customer premises equipment of the private networks. Advantageously, the network can be used to switch communication channels at switching points in the network to allow the customer to redirect or reallocate subscribed capacity among the customer's various private network locations so that that capacity may be used most advantageously. This is accomplished by selecting different channels from the dedicated pool to be used for handling that customer's most immediate traffic needs.

The DSFN is engineered so that the prespecified peak demands of all the private customer networks can be met simultaneously at any point in time. This engineering is based on limiting the access of each access point of each private customer network and limiting the set of such access points which may communicate for each private customer network. This arrangement differs from the current public switched network in that demand is unconstrained and the network is provisioned to carry a forecasted peak demand and is partially idle during off peak hours. The engineered pool is augmented by physically diverse facilities and adequate additional capacity sufficient to allow for the restoral of normal service in the event of failure of part of the regular facilities. When a failure occurs in the facility carrying channels which have been assigned to a particular private network, the failure may be detected in the customer's equipment; a request message is then automatically sent to the DSFN to reassign traffic on those channels to other available facilities. When the DSFN uses elements of a public network such as toll switches, the control of the public network can be used to control this function also. Advantageously, the fraction of additional communications channels which must be provided to assure the required level of reliability is lower in a DSFN than in a group of disjoint private networks. Advantageously, in such a DSFN, when repaired facilities are restored to service, their communications channels are automatically made available in the pool of facilities for use by other private networks by making these channels available in the data tables of the switches. This is made possible because the pool of facilities of a DSFN are subject to an overall flexible repair and administration scheme which makes a restored facility immediately available for carrying new traffic.

The invention provides for the administration of private networks as part of the overall process of administering the shared public network. Such administration takes advantage of the economies of scale offered by the large administrative systems that are present in shared public networks. Indeed, this reduces administration costs for such private networks.

The public switched network illustratively provides a CCITT standard Integrated Services Digital Network (ISDN) interface for communicating with customer access equipment. Such an arrangement permits a wide range of customer equipment to interface in a standard way with the DSEN. It also provides out-of-band signaling to provide the command and control signaling allowing the communications channels to be combined into variable transmission rate groups and allowing reestablishment of failed channels.

The invention enables two or more private networks to share units of capacity such as a 24 channel T digital carrier facility and resultingly to increase utilization of interoffice facilities.

A feature of the invention is that a customer

administrator of a private network initiates a change of facilities to respond to changes in the traffic pattern of a private network by means of the aforementioned signals from the private network. Accordingly, the private network is immediately reconfigured to meet this request, without requiring intermediate processing of service orders by the public network administrator.

It is a feature of the invention that failures of facilities transporting customer channels are signaled via the aforementioned standard out-of-band signaling channel to customer access equipment. In response the customer access equipment signals for a reconnection which is routed by the network switch over a diverse network facility with redundant capacity.

Illustratively, each private network customer has access to the public switched network via dedicated access facilities. Connections within the customer's private network provided according to this invention are established only between these dedicated access facilities. These dedicated access facilities therefore define the scope of the private network assigned to that customer. These constraints define the required traffic capacity of the shared pool of facilities and permit the shared pool to be properly provisioned. As a result, customer administrators are free to request facilities for any traffic pattern meeting these constraints, i.e., the dedicated access facilities, without causing an overload of the shared pool; effectively, if a private network administrator legitimately requests more capacity for one route, capacity of other routes is diminished.

In accordance with one aspect of the invention, a customer service controller is used for interfacing between the customer equipment and the network. The controller detects failures in the communication channels of a path and automatically sends a message to the network to request the establishment of an alternate path. The controller also maintains a record of the status of the customer's network configuration. In a preferred embodiment, the controller signals to the network over a Primary Rate Interface (PRI) of an Integrated Services Digital Network (ISDN) connection to the network.

Therefore, in accordance with the principles of this invention, a pool of transmission channels interconnecting a plurality of switching systems is dedicated for use by a plurality of private networks which request connections between their access facilities to the switching systems by sending a request message to one of the switching systems which responds to this message by causing the requested connection to be established using channels selected from the dedicated pool.

Brief Description of the Drawing

FIG. 1 is a block diagram of prior art private network arrangements;

FIG. 2 is a block diagram of private network arrangements conforming to the principles of this invention;

FIGS. 3-5 are examples of separate and total network demands of two private customers;

FIGS. 6-17 illustrate the advantage of using shared as opposed to dedicated networks;

FIG. 18 illustrates the process of setting up a connection for a private network;

FIG. 19 is a flow chart illustrating the steps of setting up such a connection; and

FIG. 20 is a block diagram of a customer's service controller.

Detailed Description

FIG. 1 is an illustration of the prior art. Customer multiplexers 90, located on customer premises are connected via carrier systems to serving offices 10 and 8. These serving offices comprise cross-connects 102 and 82 respectively, for connecting individual trunks from one carrier system to another carrier system that interconnects the serving offices to an intermediate tandem office 14. The tandem office also comprises a cross connect 142 for interconnecting trunks terminating on the tandem office. More generally, serving offices may be interconnected through a plurality of tandem offices, through other serving offices or directly, in all cases via a cross-connect facility.

FIG. 2 is an overall block diagram illustrating one embodiment of the invention. A transmission network 220 comprising 7 serving offices, 2,4,6,8,10,12 and 14 is used for interconnecting the private service customers. In this embodiment, the 7 serving offices are toll switches such as the 4 ESS™ switches described in The Bell System Technical Journal, Vol. 56, No. 7, September 1977, pages 1015-1320, and comprise switching networks for setting up temporary or long term connections and facilities switching arrangements for taking incoming groups of channels and routing them to outgoing groups of channels. The 6 serving offices on the periphery, offices 2,4,6,8,10, and 12 are each connected to one other peripheral serving office and are each connected to the central serving office 14. Each of the 7 serving offices, 2,4,6,8,10,12 and 14 is also connected to a common channel signaling network 230 which is used to pass signaling information among these switches and which is used for accessing a data base called a Network Control Point (NCP) 240. Each of the serving offices also has additional channels 1 for connecting to other serving offices; these other

channels together with the connections shown form the pool of channels dedicated for providing service to a plurality of private service customers. The Network Control Point 240 is used for translating signaling information into network physical addressing for routing and is used in conjunction with restricted access to the pool of channels to ensure that individual private networks do not exceed their assigned capacity. Connected to the NCP 240, is a service administrator's terminal 242 used for administration of the customer specific data relating to network 220 in the NCP 240, and for assigning channels to the dedicated pool.

In this embodiment, the shared facilities are derived from carrier systems common to the public telephone network but carry only connections originated from private network users who subscribe to the DSFN. Public telephone traffic is carried on separate trunking facilities. This assignment along with the demand restricted by the access facilities assures a level of availability comparable to that of dedicated facilities which is unaffected by unusual public telephone network demand. In an alternative embodiment of the invention, advanced routing algorithms can be used which allow the facilities to be further shared with public telephone traffic. Such algorithms logically reserve channels on the shared trunking for each service and allow priority, for example, to DSFN when overload conditions develop. Such routing schemes create further economies of scale in sharing but may have different performance characteristics.

Block 250 illustrates a typical private network user connection to network 220. The interface between the shared transmission network 220 and the private customer is a service controller 255 connected to a serving office 12 by a facility operating under the protocol of an Integrated Services Digital Network (ISDN) Primary Rate Interface (PRI) 257. This PRI comprises 23 B-channels, each 64-Kilobit per second (Kb/s), for carrying voice or data communications, and one 64 Kb/s D-channel for carrying signaling information. The service controller 255 is connected to an administrative terminal 260 for administering the private network by, for example, entering the network addresses of endpoints to be connected or receiving status information concerning endpoints. Since much of the traffic carried on these private networks is data traffic, a connection is shown from the service controller to a computer 264. A user terminal such as terminal 274, but at another customer location connected to another serving office, is connected via the network 220 to computer 264. The voice and switched data traffic of the private network is transmitted via connection 276 between the service controller 55 and private branch exchange (PBX) 270. A second connection via link 278 connects the service controller to the private branch exchange for serving public switched network traffic from PBX 270. This connection may share the access facility to the serving office 4 for the purpose of placing calls on the public network. The PBX is connected to user station 272 and user terminal 274.

Service controller 255 is a Unit, such as the Acculink® Model 740 of AT&T Information Systems, which can readily be adapted to perform the following functions:

1) Terminate and monitor the individual network access transmission channels 257.

2) Implement the PRI signaling protocol in order to interface with the signaling required to establish and control connections from serving office 4.

3) Implement facilities to maintain the identity of allowable destinations and allowable associated bandwidth for each of these destinations so that no attempt will be made to use channels in excess of those allocated for the private customer network. Such an attempt would be blocked by the network 220.

4) Implement transmission interfaces, including less than T1.5 rate, to other customer premises equipment to allow for switching or multiplexing onto established connections.

5) Monitor the PRI signaling protocol for connection failures. Re-establish any failed connections via PRI call setup procedures. These are specified by Consultative Committee on International Telephone and Telegraph (CCITT) standards for the ISDN PRI.

6) Implement service controller-to-service controller communications, e.g., via the PRI end-to-end signaling connection.

7) Analyze and report diagnostic data obtained via PRI signaling and report such data to administrative terminal 60. These functions are described in detail with respect to FIG. 18.

The PRI is described, for example, in "AT&T Integrated Services Digital Network (ISDN) Primary Rate-Interface Specification", AT&T Document TR 41449, March 1986, and in the CCITT Red Book Vol. 3, Fascicle 3.5, Series 1, ISDN Recommendations, Geneva, 1985.

FIGS. 3-17 illustrate how a network such as the one described with respect to FIG. 2, can be used to provide private service economically. A number of problems are solved by these arrangements. The problems include the following:

1) In prior art private networks, the use of a cross connect system to provide customers with a fraction of a carrier facility isolates the customer access equipment from existing methods of detecting facility failures. The use of the shared private network arrangement permits customers to use only a fraction of a transmission facility (e.g., 6

channels of a T-carrier facility) while receiving failure messages via the out-of-band signaling channel.

2) For many private networks, it is necessary to provide redundancy so that in case of a failure, the private customer's operation is not shut down. In many cases, the provision of an alternative route under the prior art arrangements, requires transmitting data over long multi-link connections which in a shared network, could be used much more efficiently.

3) In prior art private networks, it is inefficient to provide a customer with a fraction of a carrier facility because it is difficult to arrange that several private customers share a particular facility, especially under circumstances where redundant facilities are required. The use of a shared private network arrangement permits customers to use only a fraction of a transmission facility (e.g. 6 channels of a 24-channel T-carrier facility).

4) In a switchable shared private network arrangement, switches can be used to reconfigure a particular private customer's network in response to changes of demand, thus, avoiding a requirement of providing facilities of capacity sufficient for all different demands of one customer, or, alternatively, to reprovision capacity by manually changing cross connections within the carrier network.

5) In a shared private network, it is possible to share redundant facilities thus providing better performance in the face of trouble with lower facilities expense. For example, if a customer needs one facility and purchases two in order to have redundancy, that customer could receive better service in the face of trouble if he shared ten facilities with other customers who had an aggregate demand of seven facilities.

6) It is desirable to provide redundant facilities using geographically diverse routes so that a common trouble source such as a cable break does not remove from service a facility and its redundant facility. A shared private network arrangement provides a much larger amount of geographical diversity for redundant routes than is economical in a dedicated private network.

FIG. 3 illustrates the traffic demand of customer A. The units are in fractions of a T-carrier facility so that one quarter corresponds to six channels of 64 kilobits per second each or 384 kilobits per second. Customer A's network consists of terminations on serving offices 10,12,2,4 and 6. All of the demand by customer A is for connections between customer A's equipment connected to switch 6 and the equipment connected to serving offices 10,12,2 and 4. There is no demand for traffic among the stations terminated at serving offices 10,12,2 and 4. The traffic demand is for one quarter unit from each of the serving offices 10,12,2, and 4 to switch 6.

FIG. 4 illustrates the demand of customer B. Customer B is connected to serving offices 2,4,8, and 10 and the typical demand is for a quarter unit of traffic between 2 and 10,2 and 4,2 and 8,4 and 8, and 4 and 10. The maximum demand generated at the connections to serving offices 2 and 4 is three quarters of a unit each and the maximum demand generated at the connections to serving offices 8 and 10 is one half unit each.

FIG. 5 illustrates alternative demands of customer B. The alternative demands are that serving offices 2 and 4 can generate or terminate up to three quarters of a unit of demand each and serving offices 10 and 8 can generate up to half a unit of demand each and that any interconnection within this restraint is allowed.

The actual physical network is shown within block 20 of FIG. 2. It comprises a star wherein serving offices 2,4,6,8,10 and 12 are connected to central serving office 14 and in addition, there are facilities between serving offices 12 and 2; 4 and 8; and 6 and 10. Thus, each of the peripheral serving offices 2,4,6,8,10 and 12 have two geographically and physically diverse output links and serving office 14 has six such links.

FIG. 6 illustrates a dedicated private network for customer A having no redundancy. For the dedicated networks, each serving office has a cross-connect facility for establishing the connections required for the private networks. Blocks 501,503,505,507 and 509 represent the interface equipment such as service controller 255 connected to a cross-connect facility in one of the serving offices. Blocks 509 and 507 are represented as being connected to serving offices 6 and 10 by a single line to indicate the fact that no tandem switching takes place at these interfaces. A double line is shown connecting blocks 501,503, and 505 to serving offices 12,2, and 4 to indicate that the interface of these three serving offices performs a tandeming function, that is, a function of switching incoming circuit demand directly to outgoing facilities as well as performing the function of connecting locally generated demand to the network. For example, interface 501 switches the demand from interface 509 directly to interface 503, in addition to adding its own demand to that headed for that interface. In contrast, the tandeming function is performed at the switching network within serving offices in the shared networks of this description.. Because, in a dedicated network the facilities may not be shared among the different private customers, each of the links 511,512,513,514,515,516, and 517 is adequate for carrying a full unit of traffic though the requirements indicated in parentheses for each of these links varies from one quarter

(links 511 and 512) to one half (link 513) to three quarter (links 514 and 515) to one (links 516 and 517). Those links interconnecting two serving offices which are not directly connected in the transmission network 20, such as serving offices 10 and 12, are shown as two separate links 511 and 512 connected by a permanent connection 520 within serving office 14. Similarly, links 514 and 515 interconnecting serving offices 2 and 4 are connected by a permanent connection 522 and links 516 and 517 interconnecting serving offices 4 and 6 are connected by a permanent connection 524 within serving office 14. Notice that in order to meet customer A's demand using a dedicated network, a total of seven links are required to provide non-redundant service.

FIG. 7 illustrates a dedicated network which can meet the demands of customer B. Interface equipment 601,603,605, and 607, each comprising a service controller 255 on serving offices 2,4,8, and 10 are used by customer B for interfacing with the dedicated network. Links 611 and 612 which are connected by connection 620 in serving office 14 interconnect customer B terminations on serving offices 2 and 4. Interface 603 provides a tandeming function to switch demand from the interface 601 to interfaces 605 and 607, as well as to interface 603. Link 613 interconnects serving offices 4 and 8 directly and links 614 and 615 connected via a connection 622 within switch 14 interconnect interfaces 605 and 607. Interface 605 also provides a tandeming function to switch demand from interface 607 to interfaces 603 or 601 as well as to termination 605. While only full facilities are used for each of the links, the maximum demand traffic required on each of these links is a full unit for link 613; three quarters of a unit for links 611 and 612; and one half unit for links 614 and 615. The total number of links required to serve customer B using a dedicated network without redundancy is five.

FIG. 8 is a superposition of FIGS. 6 and 7 and indicates that to meet the total demand for the private networks of customers A and B using dedicated facilities requires 12 links: one unit of traffic over route 701 between serving offices 12 and 2; two units of traffic over the route 703 between serving offices 2 and 14; three units of traffic over route 705 connecting serving offices 14 and 4; one unit of traffic over route 707 connecting serving offices 4 and 8; one unit of traffic over route 709 connecting serving offices 8 and 14; two units of traffic over route 711 connecting serving offices 10 and 14; one unit of traffic over route 713 connecting serving offices 12 and 14; and one unit of traffic over route 715 connecting serving offices 6 and 14.

In the configurations of FIGS. 9-11, switching is performed at each serving office and partial units of traffic can be merged onto a full unit of traffic using this switching capability. The interfaces at the customer premises comprise units 255, and the interfaces are shown as blocks 502,504,506,508 and 510 for customer A, and as blocks 602,604,606 and 608 for customer B. In FIG. 9, customer A's demands can be met using routes 801,803,805, and 807 joining serving offices 12,2,4 and 10, respectively, to serving office 14 and each carrying one quarter of a unit of traffic to route 809 connecting serving office 14 to serving office 6 and carrying a full unit of traffic. This requires only five links in contrast to the seven links required in FIG. 6.

FIG. 10 illustrates how customer B's demands can be met through four links over routes 901,903,905, and 907 carrying three quarters, three quarters, one half and one half a unit of traffic apiece. This contrasts with five links required in the configuration of FIG. 7 to carry the same traffic.

Finally, FIG. 11 illustrates that six links can carry all the traffic required by customers A and B. As can be seen from examination of FIGS. 9 and 10, a single route 1001 is all that is required to carry traffic from serving office 14 to serving office 12 for termination on unit 502; a single route carrying one unit of traffic is all that is required for connecting serving office 14 to serving office 2 for switching of that traffic to interfaces 504 and 602; a single unit of traffic is all that is required over link 1005 connecting serving office 12 and serving office 14 for switching traffic to interfaces 506 and 604; a single link carrying one unit of traffic is all that is required for route 1007 between serving offices 14 and 6; interface 508; a single link is all that is required on route 1009 connecting serving office 14 with serving office 8 to carry one half unit of traffic to interface 606 and a single link is all that is required on route 1011 to carry three quarters of a unit of traffic switched between serving office 10 and serving office 14 for interfaces 510 and 608. This is half the number of links shown in FIG. 8 to meet the demand of customers A and B using dedicated facilities.

FIGS. 12-14 illustrate the configuration required to achieve a reliable dedicated network. A reliable network is defined for the purposes of these figures as a network which will survive the loss of any one link. FIG. 12 shows that a reliable network for serving customer A can be implemented using nine links. The links interconnecting switches 12,2; 2,14, and 14,4; 6,14, and 14,8, and 8, 4; 6,10; and 10,14 and 14,12. The same set of links differently configured as shown in FIG. 13 can provide customer B with a reliable dedicated network. The links include one connecting switches 2,12 and 12,14 and 14,10; 2,14 and 14,4; 4,8; 8,14 and 14,6 and 6,10. Notice that in both of these cases, several double and triple route links are required to join two serving offices. This is a consequence of the desire to

avoid using the same route for redundant links. FIG. 14 is then simply a superposition of the links of FIGS. 12 and 13 and requires a total of eighteen links.

FIGS. 15-17 demonstrate that this number can be halved by using a shared network approach. FIG. 15 illustrates that the links needed for reliable interconnection of terminations 502,504,506,508, and 510 of customer A's network is nine links, the same nine links as indicated in FIG. 12. Similarly, (FIG. 16) nine links are required to provide customer B with a reliable shared network. However, to provide both customers A and B with a reliable shared network, requires the same nine links, (FIG. 17) thus, halving the total number of links required for a reliable network using unshared facilities without switching.

In these examples, FIGS. 6 and 9 illustrate the advantage of a shared network in eliminating backhaul, i.e., the process of going through a number of serving offices in order to connect adjacent serving offices. FIGS. 9-11 illustrate the sharing of channels on a facility to reduce the total number of links required. FIGS. 7 and 10 illustrate the advantages of use of switching to respond to shifts in demand. Finally, FIGS. 12 through 17 illustrate how redundancy can be obtained at lower cost in a shared switchable network. FIG. 14 illustrates that eighteen links are required to achieve a reliable dedicated network, whereas, FIG. 17 shows that only nine links are required if the network can be shared. Finally, FIG. 17 shows a redundant and geographically diverse network wherein traffic over a facility containing any cable break can be routed over another facility.

FIG. 18 illustrates the procedures and network data necessary for the service controller to establish a connection, and the means by which the network switching elements are able to control demand and route requested connections. The customer network administrator at terminal 260 (FIG. 2) is assigned by the carrier network administrator access channel numbers and dialable addresses (directory numbers) for each of the endpoints subscribed to the network. These data are entered into the service controllers by the customer admimstrator and into the network switch and Network Control Point by the network administrator. Further the network administrator assigns a network address for each endpoint as well as a set of address translations which convert dialable addresses to network addresses. The latter translations which are customer specific are placed in the Network Control Point and associated with each network address. Translations are provided only for allowed calls. Each of the serving offices comprises a switching network for establishing connections.

Service controller 255, shown in detail in FIG. 20, comprises an interface 2003 for interfacing with the access channels to serving office 4. The5interface is a PRI interface with one signaling channel 2005 for communicating with a customer administrative terminal 260 and a processor 2000 for controlling the service controller. The processor 2000 comprises a central processing unit 2002, a memory 2009 including table 2010 for storing access patterns of the service controller, table 2006 for storing channel usage data, and a program 2007. The processor 2000 communicates with serving office 4 via the signaling channel 2005. The output of the interface is also connected via channels 2004 to the customer equipment 264,270.

A call is illustrated between dialable addresses W on serving office 4, and X on serving office 8; both W and X are dialable addresses of customer B. (Dialable addresses are upper case letters W, X, Y, Z; network (physical) addresses are lower case letters w, x, y, z.) The service controller 255 upon command from the customer administrator's terminal 260, checks in Table 2010 to see how many channels can be accessed for the various allowable outgoing addresses, X, Y, and Z. Segment 2014 indicates that 12 channels are allowed for X, 12 for Y, and 18 for Z. Segment 2010 indicates that the total number of channels allowed at any one time is 18, identifies these as 1-18, and identifies the directory number, W, of the source. Service controller 255 Initiates a call to X by sending a setup message 2020 over the primary rate interface (PRI) containing the following information elements: dialable addresses X and W (2021), number of channels (6) (2023), assigned channels (7-12) (2025) and the type of service (DSFN) (2027) for which the connection will be established. The serving office 4 interprets the service type and network address in order to check a table 2010 of allowed access channels. If the requested channels are correct and adequate for the desired call, the serving office 4 formulates a Common Channel Signaling (CCS) message (2040) containing the network address w of the originating access line and dialable address X, (segment 2042), service type (segment 2044), and a call ID (segment 2046). This message is transmitted to the Network Control Point (NCP) 240. The network address along with the service type determines the message routing to the correct NCP and customer data record in the NCP. If the number or type of access channels is not correct, or is inadequate, the call is blocked by the switch by signaling a call denial message to the source controller.

The Network Control Point 240 uses the received information to verify that the customer has subscribed to the desired service, and to access a table 2050 containing the network address translations, i.e., translations from the dialable address to

the network address, for allowed network addresses for calls originating from network address w. The network address is a concatenation of the terminating switch address and the terminating address of the access line at that switch. If the desired dialable address, X, is an allowed translation for w, then a return message 2060 containing the call ID 2064 and the routing translation (2062) comprising the network addresses w, x, of the two endpoints is sent to the originating serving office. If not, a call denial message is sent to the serving office which in turn blocks the call. The details of the call denial are returned to the customer administrator and provide useful trouble shooting information.

In the case of an allowed call, the data at the serving office associated with the call ID and the translated routing number are used to proceed with the routing of the call. Service type, network address and bandwidth data are used to point to a routing table 2070 within the serving office 4 which contains a list in preferred order of the outgoing facilities which can transport the call to its destination. Only the terminating switch address portion of the network address is needed to determine routing. These facilities are dedicated to the service and engineered to give a very low level of blocking given the known demand and statistical failure rate of facilities. The set of outgoing facilities traverse at least two physically diverse paths. The preferred routes are searched in order until an idle set of channels is found. In this case, the two routes, in order of preference, are trunk group m (2082), a direct route to destination serving office 8, and trunk group n (2084), a route that goes via tandem serving office 14. Assume that a trunk of trunk group m is available. A table (not shown) of trunks of trunk group m is searched to find an available set of channels for the connection.

Subsequent serving offices in the network, based on incoming trunk type and Common Channel Signaling data in the form of call type and terminating network address will further route the call to the desired terminating switch. At any point where the call cannot be further routed due to lack of facilities, the call will be terminated. Routing schemes such as Dynamic Non-Hierarchical Routing which allows the call to be automatically reattempted at the originating switch until all possibilities are exhausted can also be employed.

In this case, with the direct route available, serving office 4 sends a CCS message 2090, comprising segments 2091 (x, network address of the destination, W, dialable address of the source), 2093(6, the number of channels for this call), 2095-(13-18, the channels of the trunk group, assumed in this case to be a single 24 channel group, that have been assigned to this call) 2097 (m, the identity of the trunk group) and 2099 (DSFN, the type of service for this call). When message 2090 is received in serving office 8, the access privilege for network address x are checked in table 2100, and it is found that the service controller 256 has been assigned to channels 1-12 on the access facility. Serving office 8 sends a message 2110 to service controller 256 to establish the requested connections. Message 2110 comprises segments 2111 (dialable addresses X, W of source and destination), 2113(6, the number of channels of the connection), 2115 (1-6, the channels assigned to the connection) and 2117 (DSPN, the type of service). Upon positive acknowledgment, the call is completed and a connect message is transmitted to the originating service office 4 and service controller 255. If the appropriate channels are inadequate in number or non-existent the call is blocked. During the progress of the call, messages relating the status of the call are sent to service controller 255. If the call is blocked, an abnormal event, a code in the denial message to service controller 255 will indicate the cause of blocking. This information would be displayed at the customer administrator's terminal 260 by the service controller 255.

Once established, a connection is held indefinitely. Severe noise or failure of the network facilities can cause the serving offices 4 or 8 or the service controller 255 or 256 to disconnect. Under these condition, the service controller 255 or 256 would automatically reestablish the connection as described above. A serving office will take very noisy or failed trunks out of service, for example, trunk groupm. When the call is reestablished, since the old route is unavailable, a new route will be selected by the serving office, in this case, trunk group n. The engineering of the network facilities assures that there will be a physically diverse facility with adequate capacity to carry the redirected connection.

FIG. 19 is a flow diagram of the functions performed as described in FIG. 18. First, a serving office receives a request to establish a DSFN connection (action block 2200). A connection is then set up between the originating serving office and the service controller from which the request was received (action block 2210). An NCP (in this case NCP 240) is then accessed to derive the physical address of the destination from the directory number of that destination (action block 2220). The originating serving process then selects a route based on the identity of the destination serving office (action block 2230). Finally, a connection is set up between the originating serving office and the destination serving office and between the destination serving office and the serving office controller at that destination (action block 2240).

For purposes of rearranging the customer network to meet changed demand, the customer administrator may disconnect existing connections and establish new connections of different bandwidths to the set of predefined locations identified in the Network Control Point. The only constraint is that the total bandwidth between two endpoints cannot exceed the bandwidth of the access channels assigned to the service. If the customer administrator desires additional bandwidth or new locations, a service order to the network administrator is necessary. If adequate network capacity exists without adding new facilities, the network administrator may grant service by updating the appropriate tables in the switch and Network Control Point. This ability to do this allows the potential of granting service in a significantly shorter interval than that required to provision a customer dedicated facility which must be manually routed.

The use of the Network Control Point (database processing system) for translating addresses allows more sophisticated features to be associated with call routing. The ability to form closed calling groups has already been described. Additional features include: translations which vary by time of day and date, subscriber activated translations changes, and centralized monitoring of connection patterns. These features could be implemented at the service controller but would be more difficult to administer and control in those devices.

It is to be understood that the above description is only of one preferred embodiment of the invention. Numerous other arrangements may be devised by one skilled in the art without departing from the spirit and scope of the invention. The invention is thus limited only as defined in the accompanying claims.

Claims

1. In a communications network comprising a plurality of serving offices and a plurality of transmission facilities for interconnecting ones of said serving offices, each facility comprising at least one communication channel, a method of interconnecting access links connecting said network to a private network customer for providing private service, comprising the steps of:
responsive to a request message from said private network customer specifying communication capacity requirements and comprising data for identifying first access links and second access links to be connected, selecting channels from a pool of channels of said transmission facilities, said pool dedicated to providing private network service to a plurality of private network customers, identities of channels of said pool being stored in data tables of said network, for providing a number of channels between said first and second access links of said customer meeting said capacity requirements; and
connecting said first and second access links via said selected channels via a switched connection through ones of said plurality of serving offices.

2. The method of claim 1 further comprising the steps of:
detecting a failure in at least one of said channels used for interconnecting said first and second access links of said customer;
selecting at least one alternate available channel from said pool of dedicated channels; and
connecting said first and second access links of said customer via said at least one alternate available channel via a switched connection through ones of said plurality of serving offices.

3. The method of claim 1, wherein said facilities comprise carrier groups each comprising a set of channels, and wherein said selecting step comprises the step of:
selecting a proper subset of the set of channels of one of said carrier group for meeting transmission requirements of said private customer between said first and second access links.

4. The method of claim 1 further comprising the step of:
communicating over at least one of said first and second access links by integrated communication and signaling channels.

5. The method of claim 1, wherein said communications network further comprises data tables for storing first data identifying the number, identity, and availability of access channels on access links of said customer, and wherein said selecting step comprises the step of:
responsive to said request, checking said request against said first data identifying the number and availability of access channels on said first and second access links.

6. The method of claim 5 further comprising the step of notifying said private customer if insufficient access channels are available on at least one of said first and second access links to meet said request.

7. The method of claim 5 further comprising the step of:
responsive to said connecting step, updating said first data to indicate that said requested number of channels on said first and second access links are unavailable.

8. The method of claim 5 further comprising the step of:
responsive to a request from a network administrator, altering said first data for identifying the number and identity of access channels on said first and second access links.

9. The method of claim 1 wherein said commu-

nications network is a common carrier network further comprising the step of:
responsive to a request from a network administrator, adding or subtracting channels to or from said pool by altering said data tables for storing identities of channels of said pool dedicated to providing private network service; and
making any channels added to or subtracted from said pool unavailable or available, respectively, for use for public service in said common carrier network.

10. The method of claim 1 wherein said communications network is a common carrier network and wherein said connecting step comprises:
connecting said first and second access links over said channels selected for said customer via a switched connection through ones of said plurality of serving offices, and wherein said ones of said plurality of serving offices comprise serving offices for also switching public telecommunications traffic.

11. The method of claim 1 further comprising the step of:
engineering said pool to meet peak demands of a plurality of private customers, said peak demands being limited by pluralities of access links provided to each of said plurality of private customers and the number of access channels provided on each of said pluralities of access links.

12. The method of claim 11 wherein said engineering step comprises:
engineering said pool to meet said peak demands augmented by adequate additional capacity to allow for restoral of channels to private customers in the event of failure of any transmission facility interconnecting ones of said plurality of serving offices.

13. The method of claim 12 wherein said engineering comprises:
engineering said pool to provide a sufficient number of channels on physically diverse facilities so that alternate channels can be provided if all facilities of one physical route fail.

14 The method of claim 11, wherein said communications network is a common carrier network, wherein said facilities comprise carrier groups each comprising a set of channels, and wherein said engineering step comprises:
engineering said pool according to rules such that said pool may comprise a proper subset of channels of a carrier group between two of said serving offices.

15. The method of claim 1 further comprising the step of:
transmitting said request from a customer administrator terminal to one of said serving offices.

16. The method of claim 15 wherein said transmitting step comprises the step of:
transmitting said request over an integrated communication and signaling facility of an access link connecting said private network customer to said communications network.

17. The method of claim 1 wherein said selecting step comprises the step of:
transmitting from one of said serving offices a message, comprising said data for identifying said first and second access links, from said request, to a data base and receiving a message in said serving office from said data base comprising data identifying at least one of said first and second access links.

18. The method of claim 17 further comprising the step of:
translating said data identifying at least one of said first and second access links to find a route from said one of said serving offices to a serving office connected to one of said first and second links.

19. The method of claim 18 further comprising the step of:
sending a message over a common channel signaling system to a serving office connected to one of said channels of said route for establishing part of said connection between said identified first and second access links of said customer.

20. The method of claim 17 further comprising the step of:
in said data, translating from said data for identifying said first and second access links to said data identifying at least one of said first and second access links.

21. The method of claim 20 wherein said data for identifying said first and second access links comprises at least one dialable number, and said data identifying at least one of said first and second access links comprises data identifying at least one serving office and a physical access channel of said at least one serving office.

22. The method of claim 1 further comprising the steps of:
storing availability data for each of said channels of said pool in said data tables;
responsive to another request from said private network customer, specifying a number of channels to be disconnected between identified access links;
disconnecting said specified number of channels; and
making said disconnected channels available, in said availability data of said data tables, for serving others of said plurality of private network customers.

23. The method of claim 1 wherein said selecting step comprises the steps of:
receiving at a first serving office said request wherein said information for identifying said access links comprises a dialable address of said second access link;
translating said dialable address of said second

access link to a physical address, said physical address comprising an identification of a second serving office connected to said second access link;
selecting at said first serving office a trunk group for connection to said second serving office; and
selecting a channel of said trunk group from members of said pool of channels available in said first serving office.

24. The method of claim 23 wherein said connecting step comprises the steps of:
setting up a first connection in said first serving office between said first access link and said selected channel; and
transmitting a message to a serving office connected to said selected channel for extending said connection to said serving office connected to said selected channel.

25. The method of claim 1 wherein ones of said channels of said pool are identified in data tables of serving offices connected to said ones of said channels.

26. The method of claim 1 wherein said connecting step comprises:
selecting a preferred trunk group comprising channels from said pool;
testing whether channels from said pool are available in said preferred trunk group; and
if no channels from said pool are available in said preferred trunk group, selecting an alternate trunk group comprising channels from said pool.

27. The method of claim 26 wherein said alternate trunk group uses different facilities than the facilities of said preferred trunk group.

28. A customer service controller for interfacing between customer equipment and a telecommunications network comprising:
means for interfacing access communications channels and signaling channels to said network and to said customer equipment;
means for detecting a failure of a communications channel of said network connected to said customer equipment; and
means responsive to said detecting for sending a request message from said controller to said network over one of said signaling channels for selecting an alternate communications channel and establishing a connection between said customer equipment and said alternate channel.

29. The controller of claim 28 further comprising:
means for maintaining the identity of allowable destinations, an allowable number of channels to each of said destinations, and present usage of said access communications channels.

30. The controller of claim 28 wherein said means for interfacing comprise at least one primary rate interface of an integrated services digital network.

31. The controller of claim 28 further comprising:
means responsive to said detection for reporting said failure to an administrative terminal of said customer equipment.

32. In a communications network comprising a plurality of serving offices and a plurality of transmission facilities for interconnecting said serving offices, each facility comprising at least one communication channel, apparatus for providing private service interconnections between first access links and second access links connecting said network to a private network customer, said first access links connected to a first one of said serving offices, comprising:
a pool of communications channels of said facilities for interconnecting said first serving office to others of said serving offices;
said first serving office operative under the control of a program for establishing connections in said first serving office for extending connections from ones of said first access links toward ones of said second access links via ones of said pool of communications channels in response to receipt of a request from said private network customer comprising data for identifying said second access links.

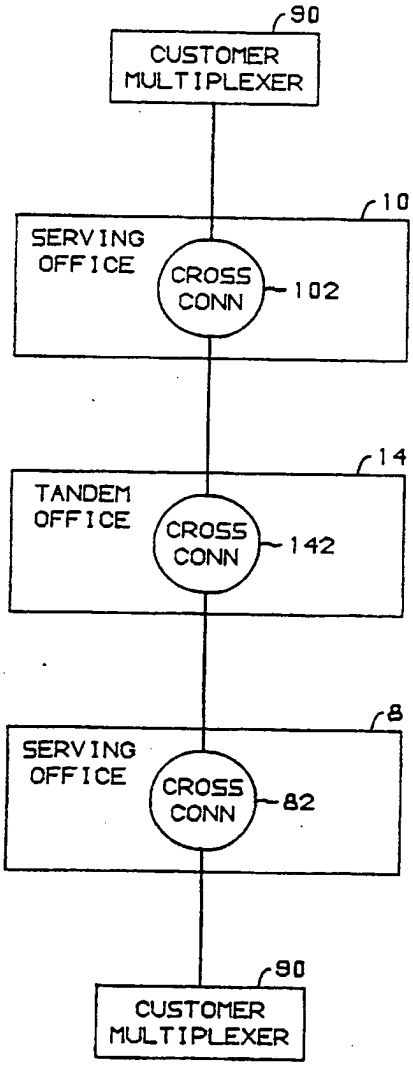
33. In the communications network of claim 32, the apparatus further comprising:
said first serving office further operative under the control of a program and responsive to a failure in one of said ones of said pool of communications channels for selecting at least one alternate available channel from said pool of dedicated channels and extending connections from said ones of said first access links via said at least one alternate available channel toward said ones of said second access links.

34. The communications network of claim 32 wherein said apparatus further comprises:
a data base system, accessible from said first serving office, for storing data for identifying the physical locations of access links of said private customer.

35. In the communications network of claim 32, the apparatus further comprising:
common channel signaling facilities interconnecting said first serving office to others of said plurality of serving offices for transmitting signaling messages for establishing connections in said others of said serving offices.

36. In a common carrier communications network comprising a plurality of serving offices and a plurality of transmission facilities for interconnecting ones of said serving offices, each facility comprising at least one carrier group, each carrier group comprising a set of channels, a method of

interconnecting access links connecting said network to a private network customer for providing private service comprising the steps of:
responsive to a request received from a customer administrator terminal of said private network customer, said request specifying a number of channels and comprising data for identifying first and second access links to be connected by said number of channels, selecting channels from a pool of channels, dedicated to providing private network service, whose identities are stored in data tables of said network, for providing said number of channels between said first and said second access links of said customer;
connecting said first and second access links via said channels selected for said customer via a switched connection through ones of said plurality of serving offices;
updating first data to indicate that said requested number of channels on said first and second access links are unavailable;
detecting a failure in at least one of said channels used for interconnecting said first and second access links of said customer;
selecting at least one alternate available channel from said pool of dedicated channels;
connecting said first and second access links of said customer via said at least one alternate available channel via a switched connection through ones of said plurality of serving offices;
communicating over at least one of said first and second access links by integrated communication and signaling channels;
responsive to a request from a network administrator, adding or subtracting channels to or from said pool by altering said data tables for storing identities of channels of said pool dedicated to providing private network service;
making any channels added to or subtracted from said pool unavailable or available, respectively, for use for public service in said common carrier network;
wherein said communications network comprises data tables for storing first data identifying the number, identity, and availability of access channels on access links of said customer;
wherein said step of selecting channels from said pool comprises the steps of:
receiving at a first serving office said request wherein said information for identifying said access links comprises a dialable address of said second access link;
checking said request against said first data identifying the number and availability of access channels on said first and second access links;
notifying said private customer if insufficient access channels are available on at least one of said first and second access links to meet said request;
in a data base of said network, translating said dialable address of said second access link to a physical address, said physical address comprising an identification of a second serving office connected to said second access link;
selecting at said first serving office, a route to said second serving office;
selecting a channel of said route from members of said pool of channels available in said first serving office; and
sending a message over a common channel signaling system to a serving office connected to one of said channels of said route for establishing part of said connection between said identified first and second access links of said customers.



(PRIOR ART)

FIG. 1

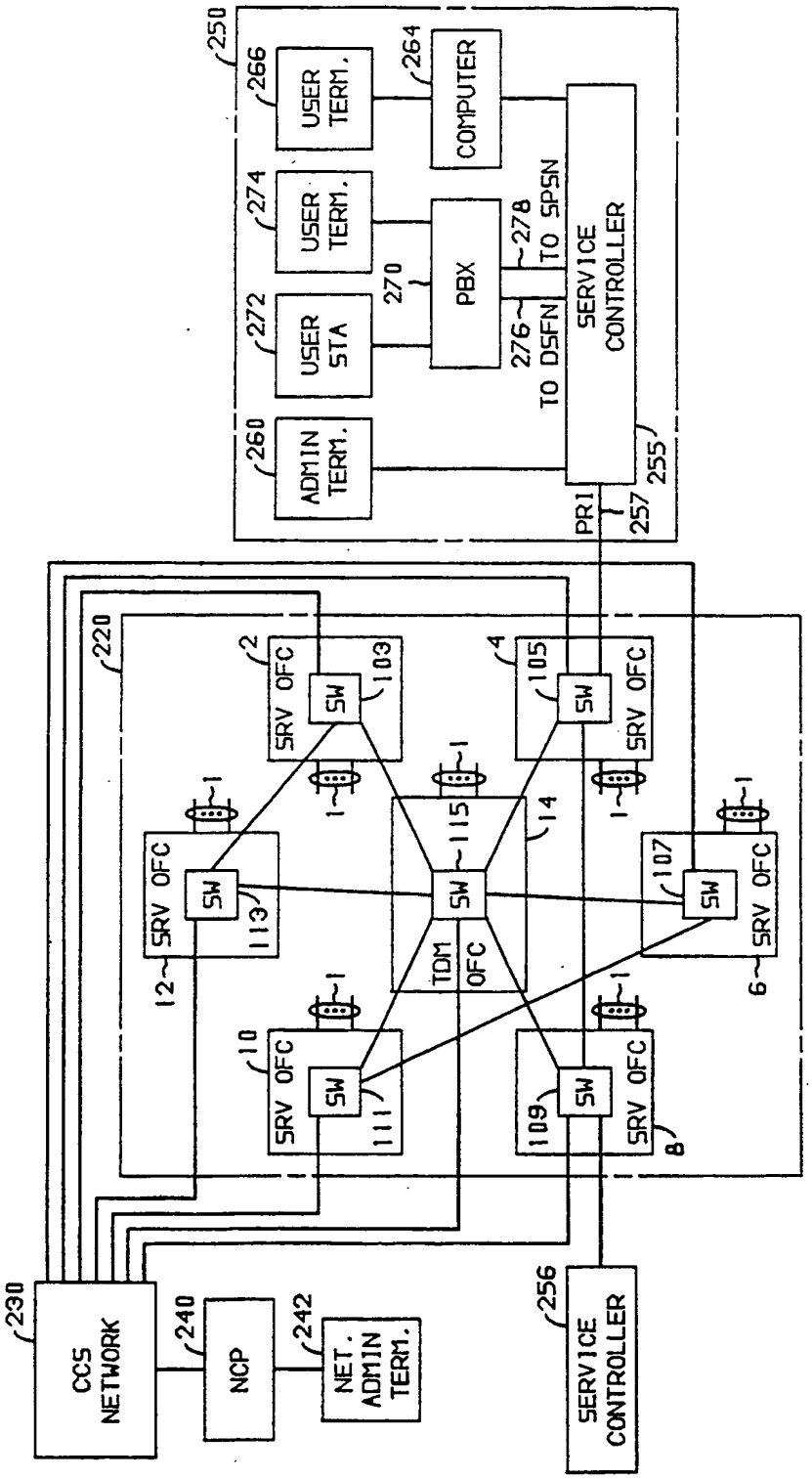
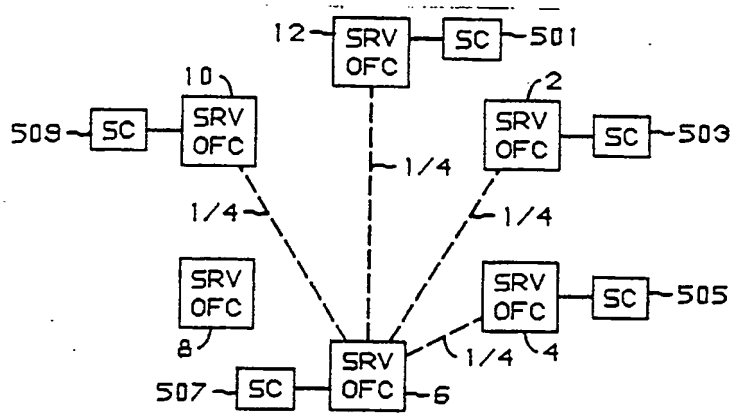
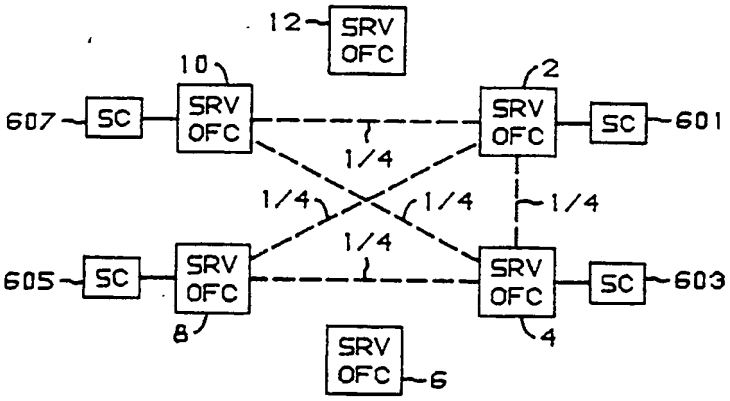

250
266
USER TERM.
264
COMPUTER
274
USER TERM.
278
TO SPSN
270
PBX
272
USER STA
276
TO DSFN
SERVICE CONTROLLER
260
ADMIN TERM.
PRI
257
255
220
2
SRV OFC
SW
103
4
105
SW
SRV OFC
1
115
14
TDM OFC
12
SRV OFC
SW
113
107
SW
SRV OFC
6
10
SRV OFC
SW
111
109
SW
SRV OFC
8
230
CCS NETWORK
240
NCP
242
NET. ADMIN TERM.
256
SERVICE CONTROLLER

FIG. 2



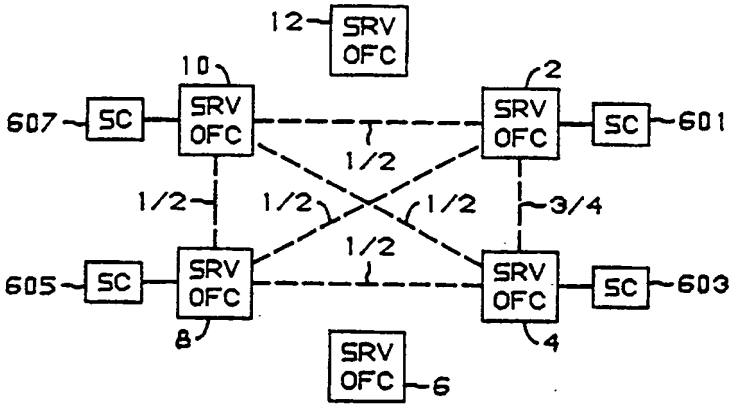
CUSTOMER A DEMAND

FIG. 3



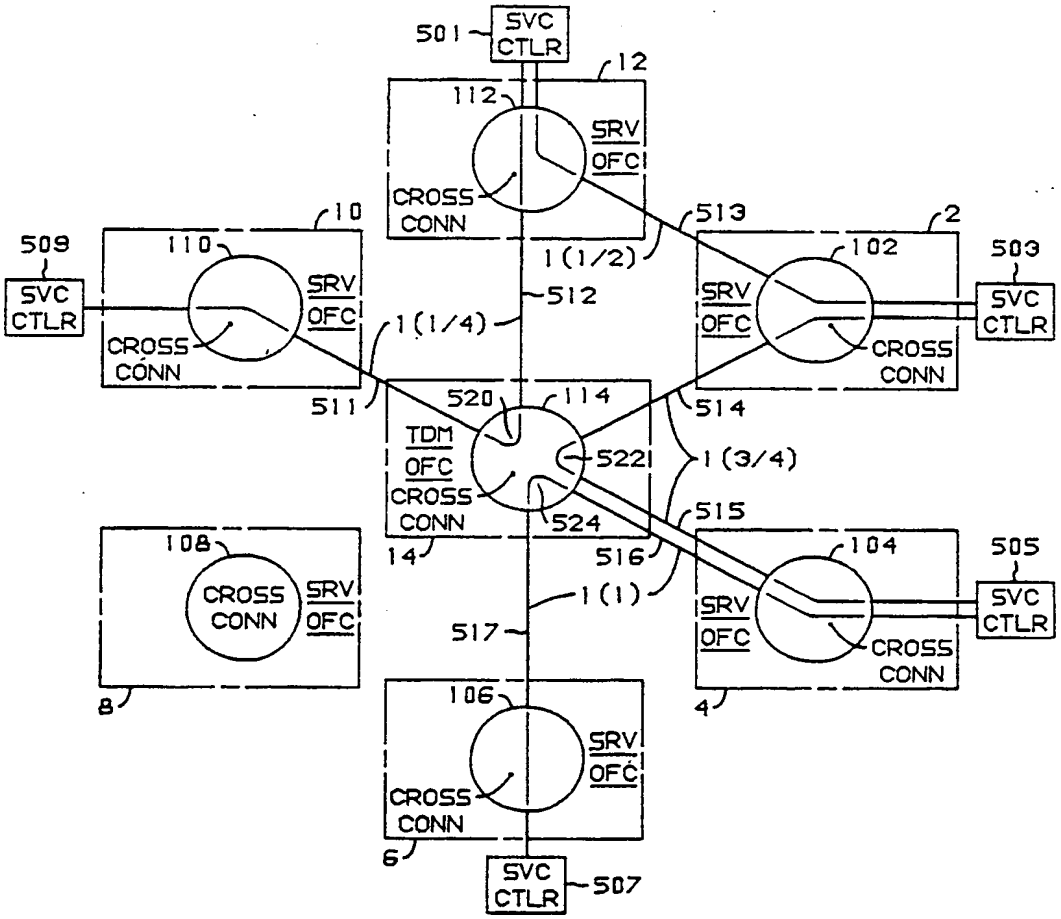
CUSTOMER B DEMAND

FIG. 4



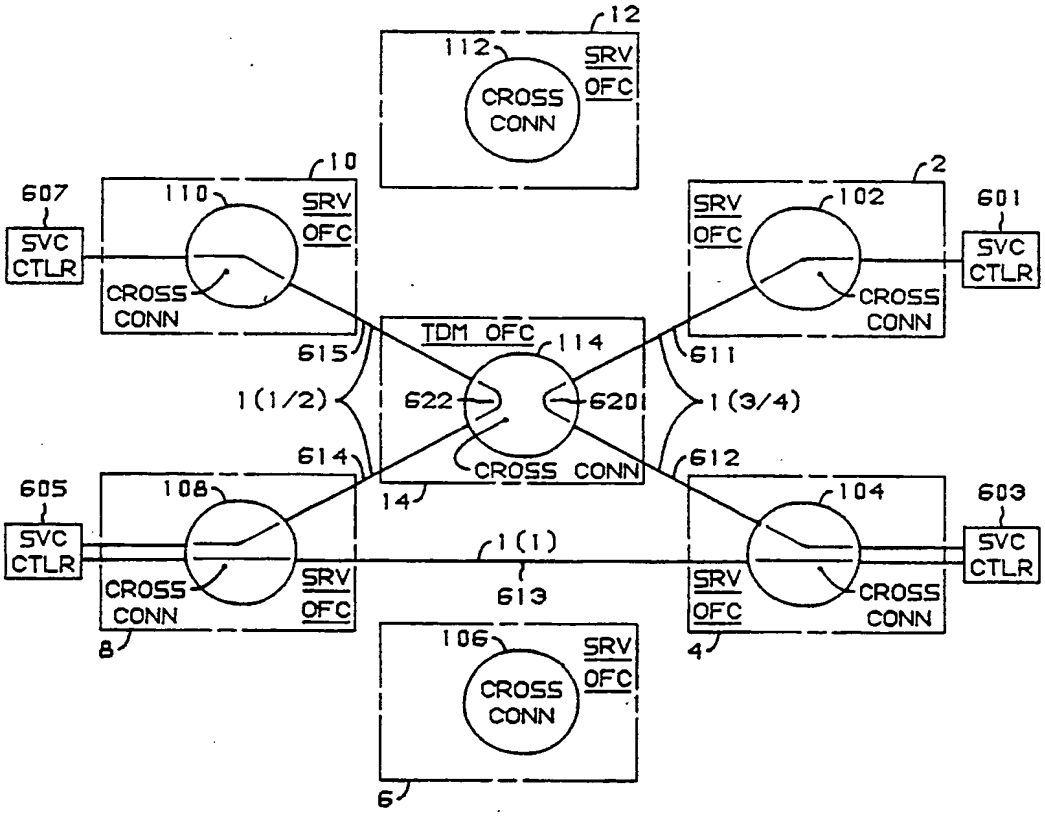
CUSTOMER B ALTERNATIVE DEMANDS

FIG. 5



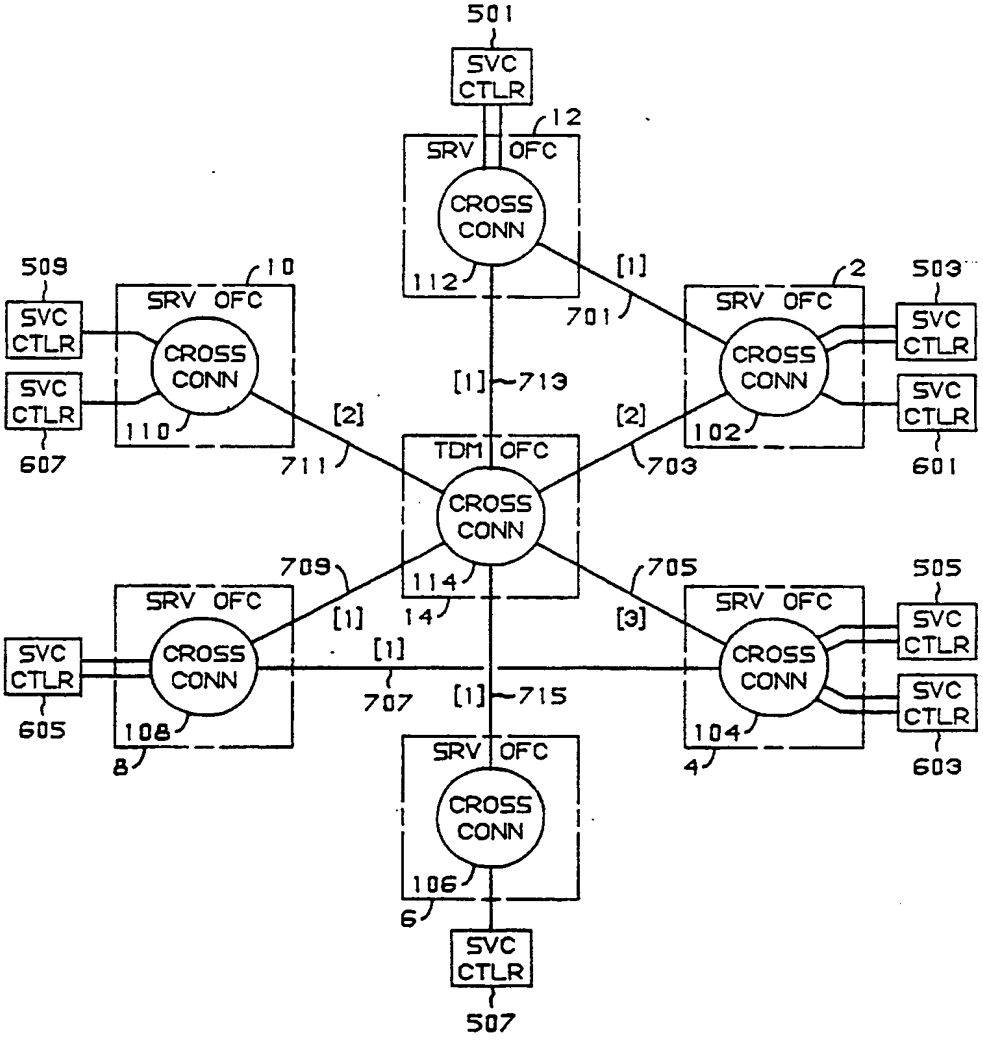
CUSTOMER A DEDICATED NETWORK

FIG. 6



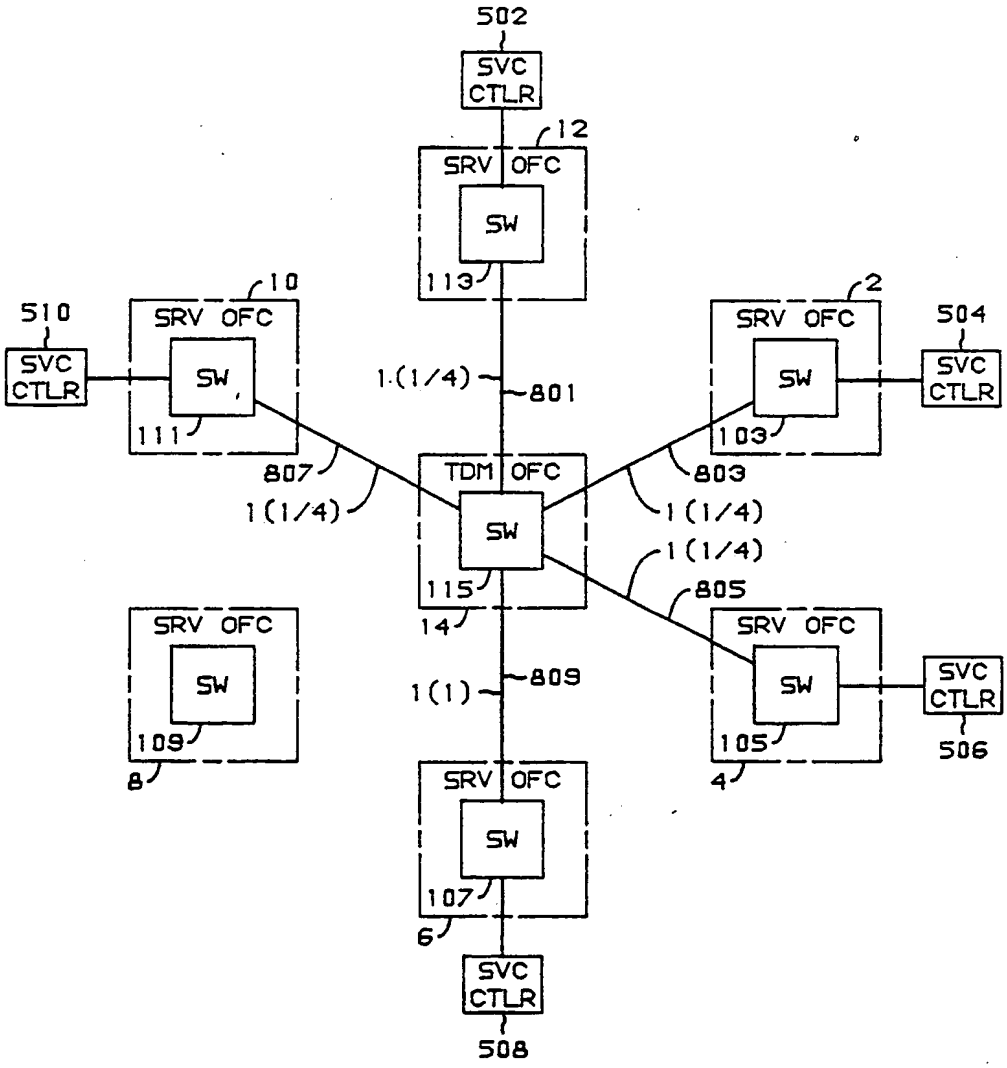
CUSTOMER B DEDICATED NETWORK

FIG. 7



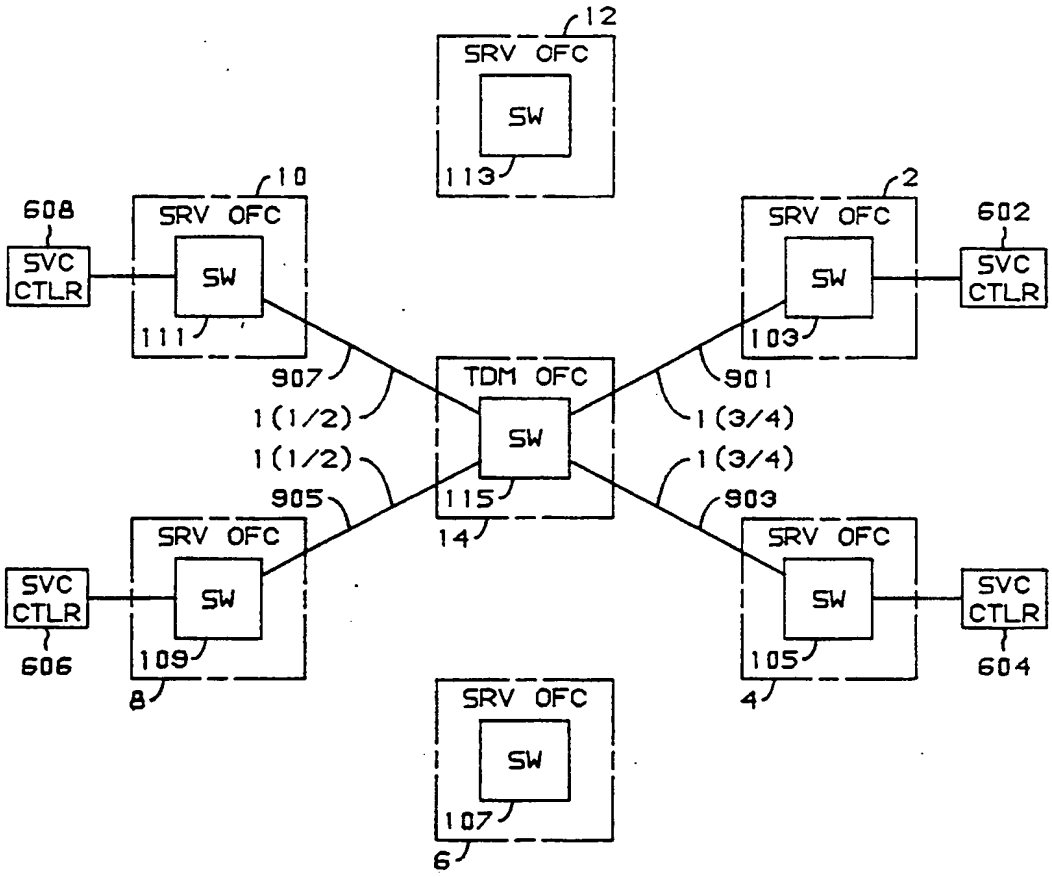
TOTAL (A+B) DEDICATED NETWORK DEMAND

FIG. 8



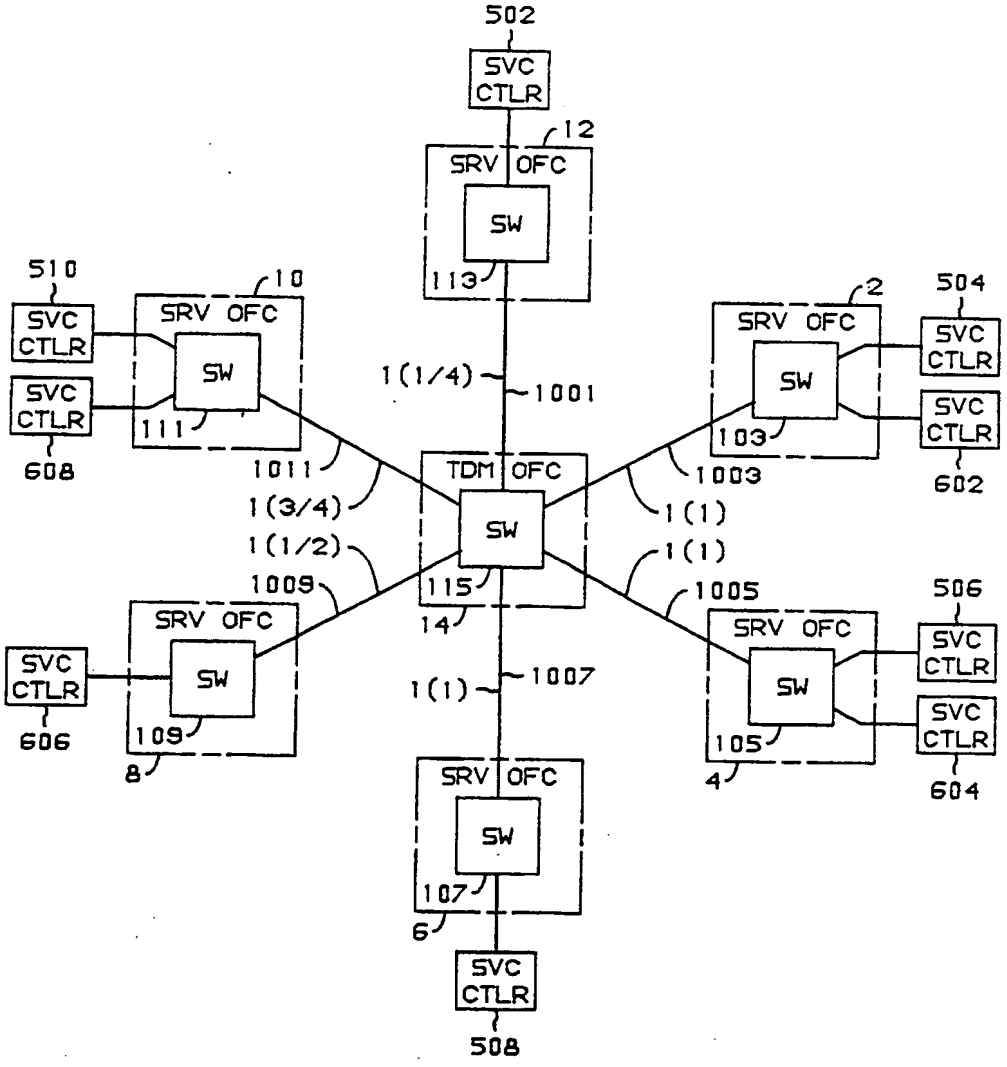
CUSTOMER A SHARED NETWORK DEMAND

FIG. 9



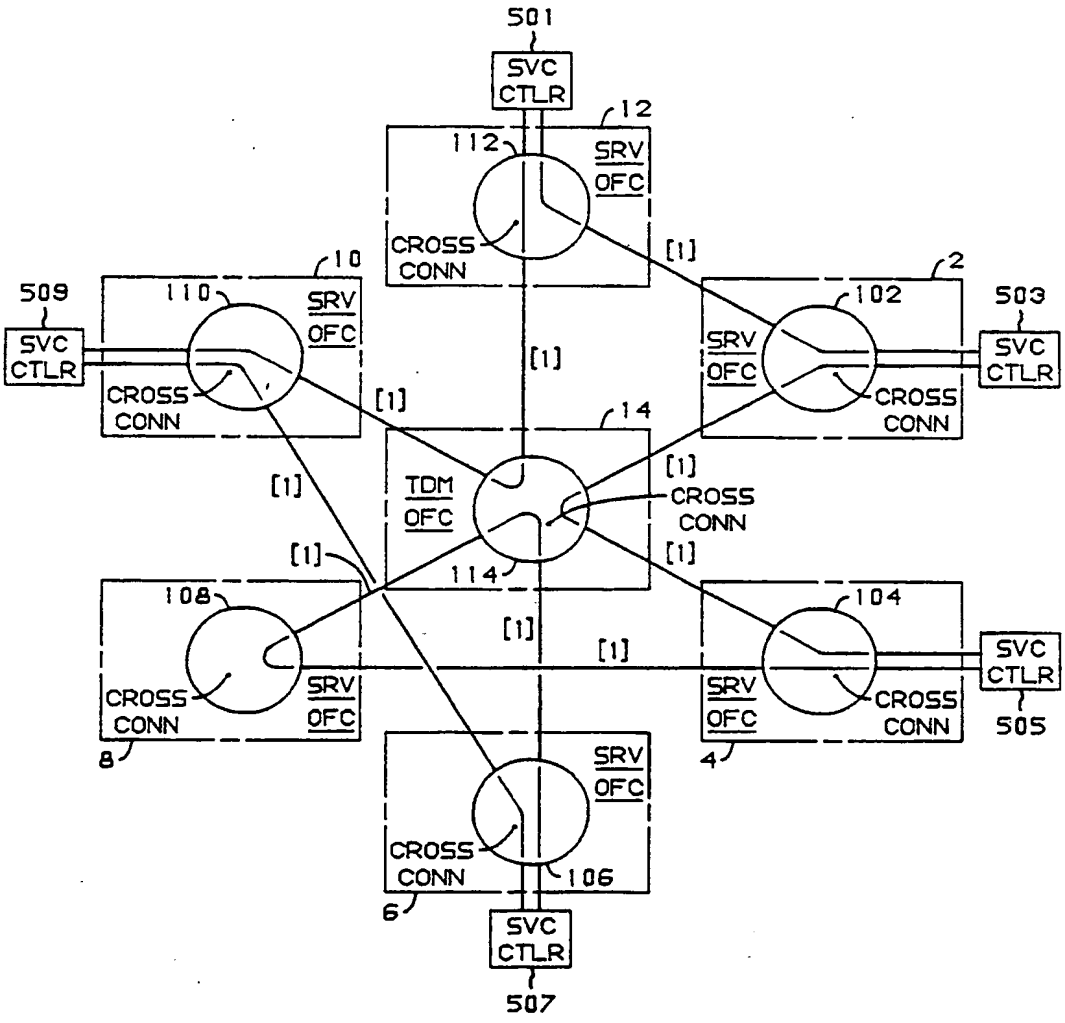
CUSTOMER B SHARED NETWORK DEMAND

FIG. 10



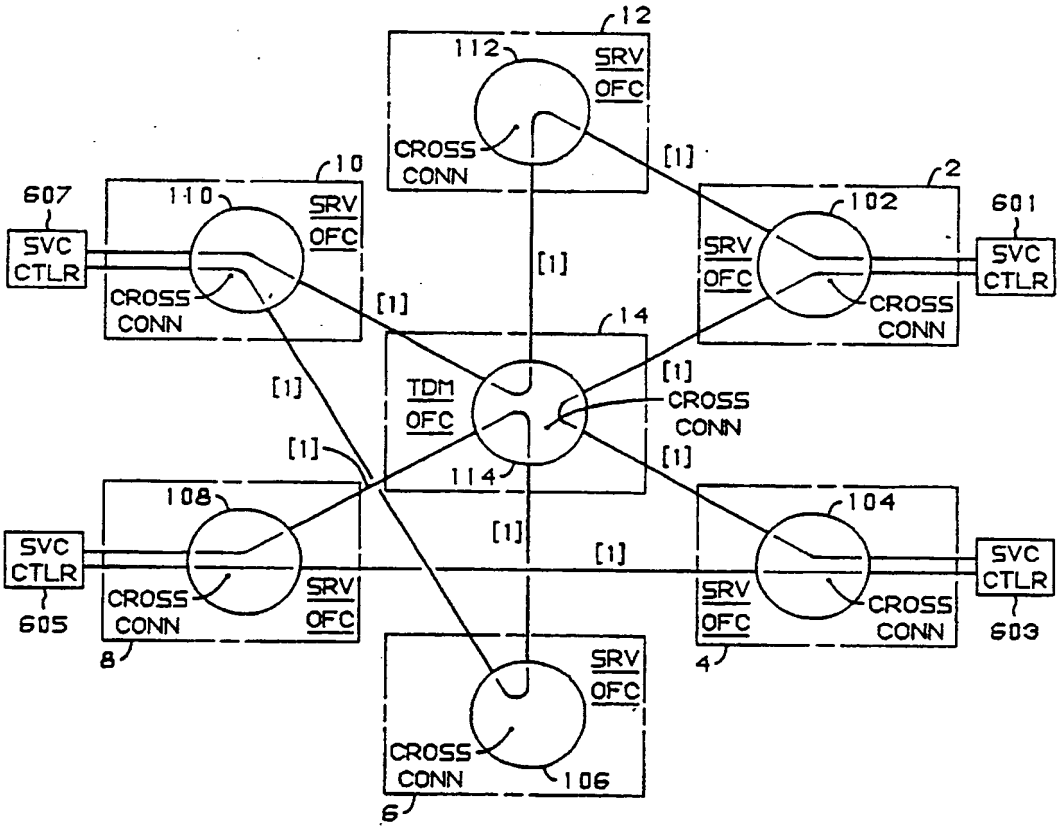
TOTAL (A+B) SHARED NETWORK DEMAND

FIG. 11



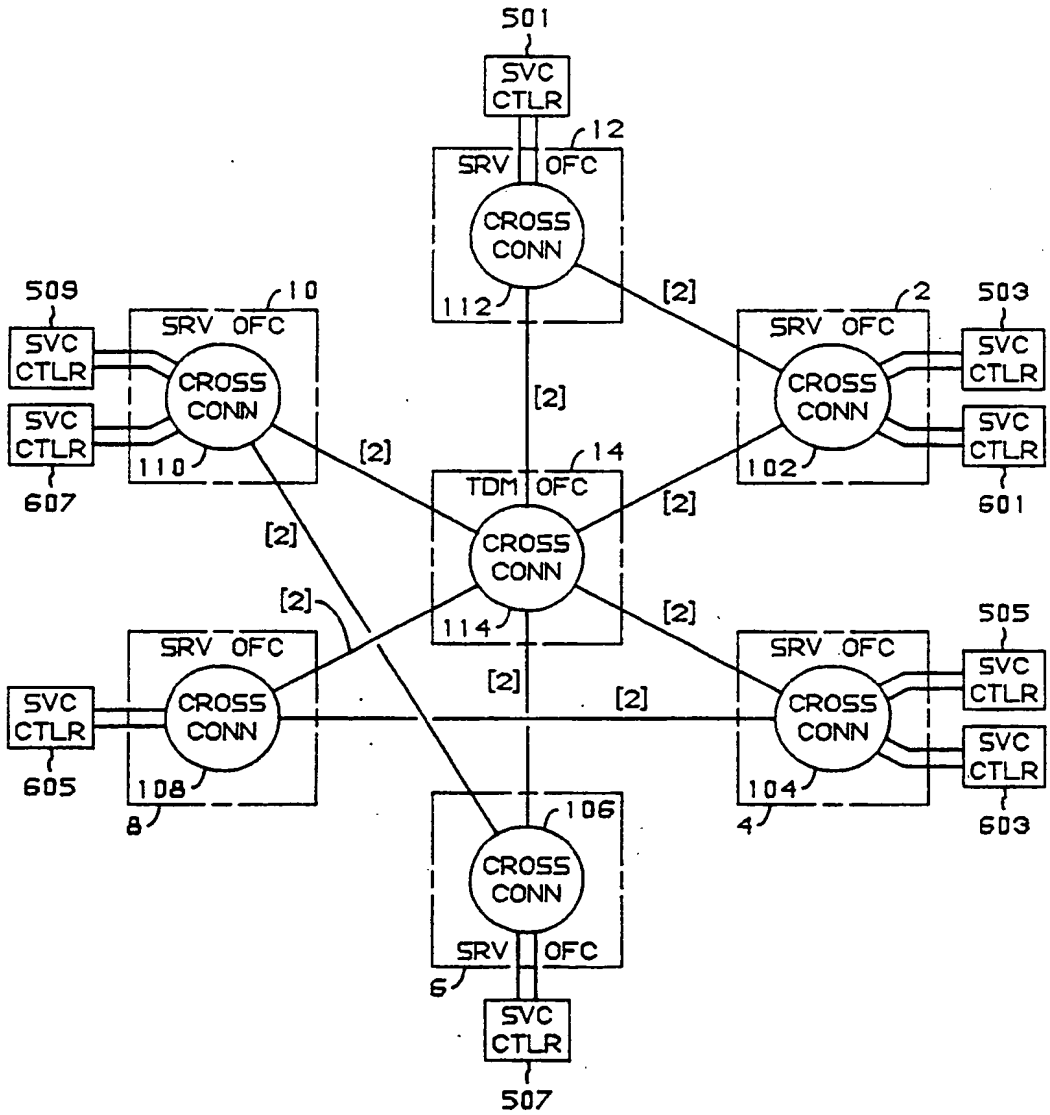
CUSTOMER A RELIABLE DEDICATED NETWORK

FIG. 12



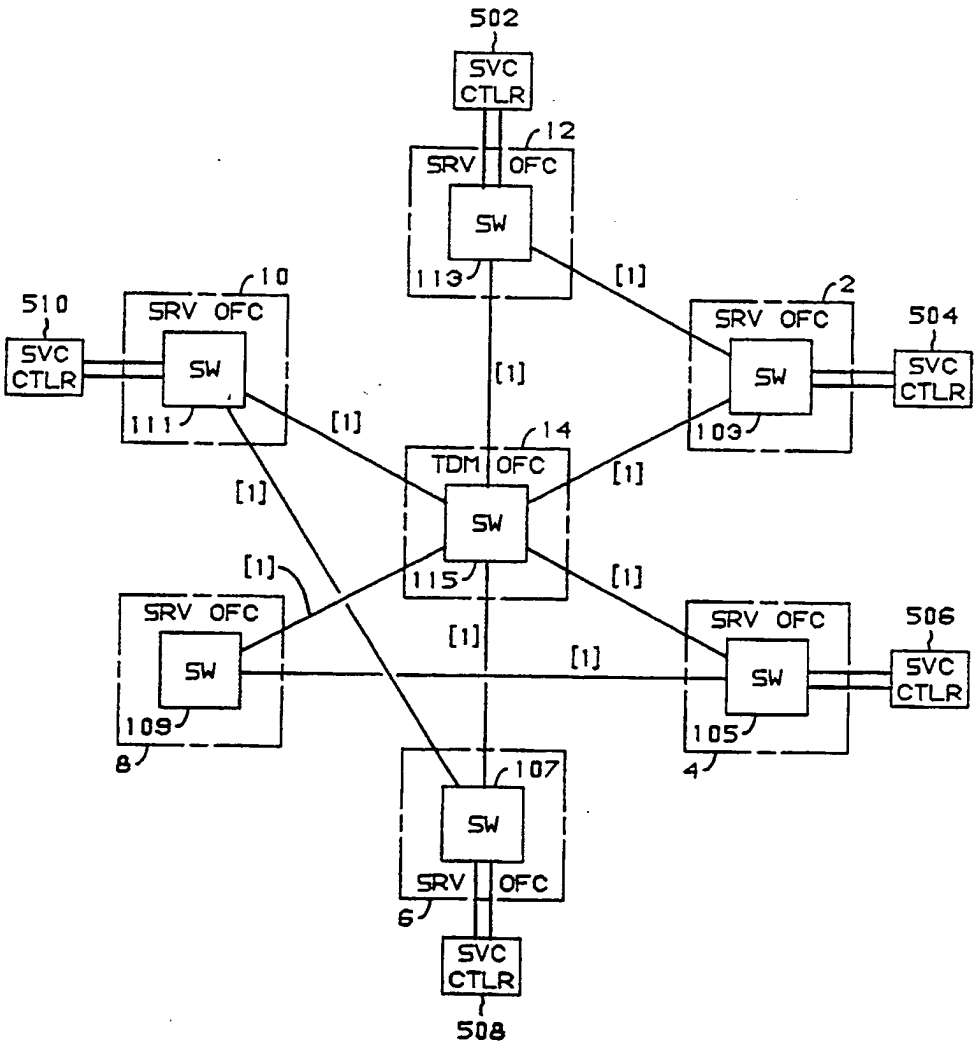
CUSTOMER B RELIABLE DEDICATED NETWORK

FIG. 13



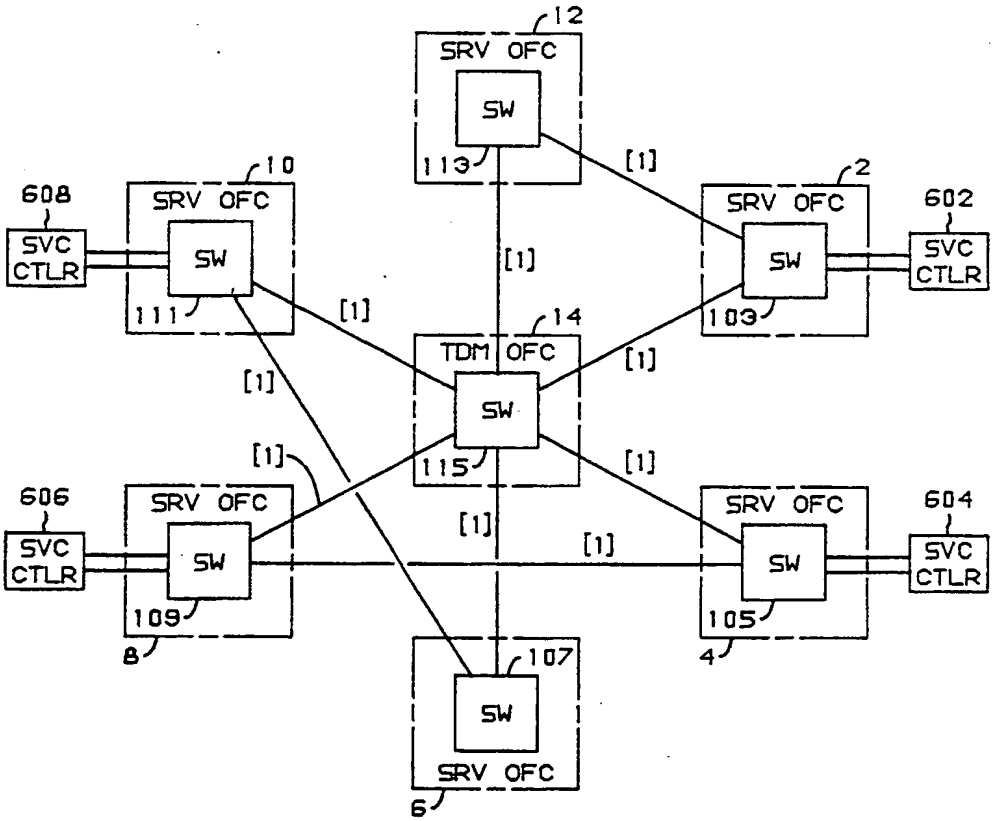
TOTAL (A+B) RELIABLE DEDICATED NETWORK DEMAND

FIG. 14



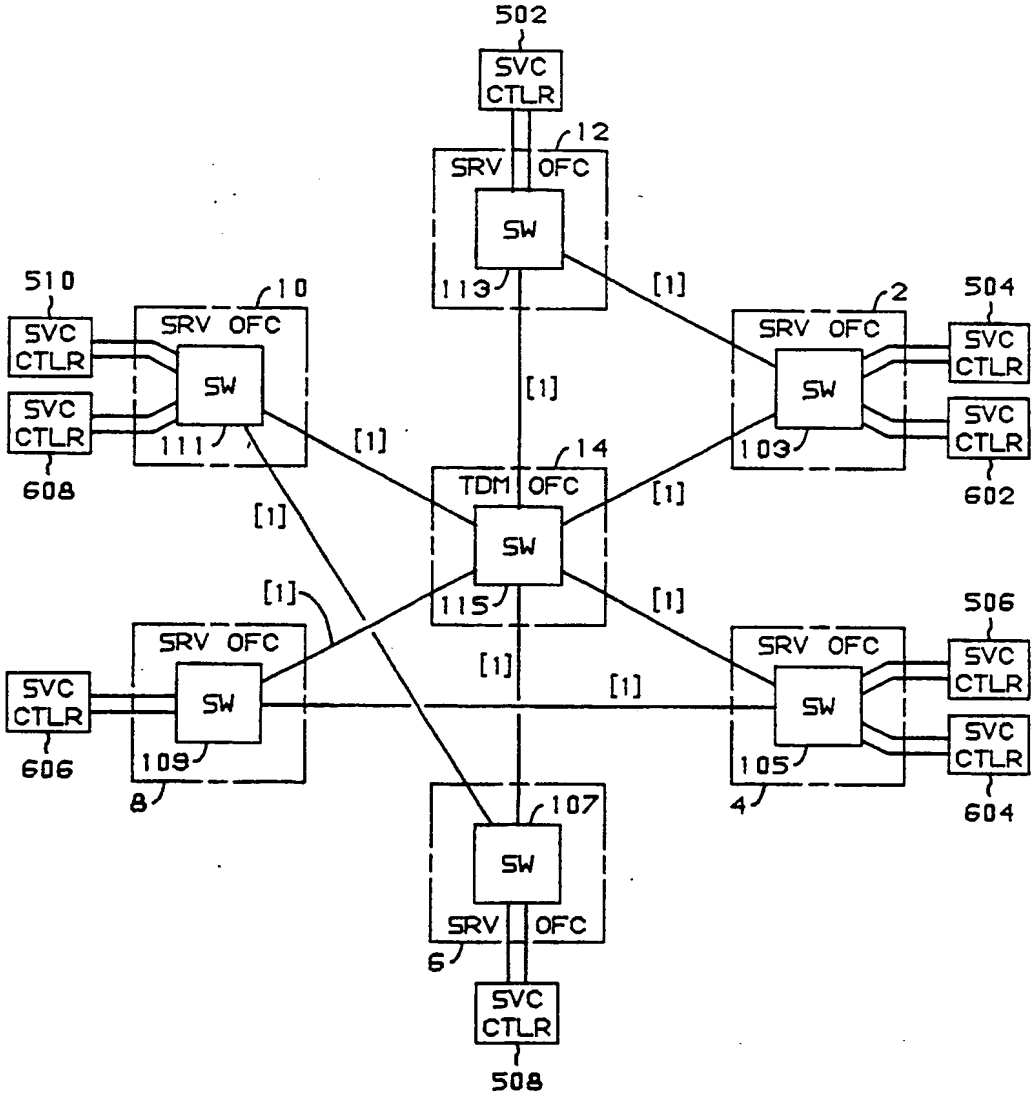
CUSTOMER A RELIABLE SHARED NETWORK DEMAND

FIG. 15



CUSTOMER B RELIABLE SHARED NETWORK DEMAND

FIG. 16



TOTAL (A+B) RELIABLE SHARED NETWORK DEMAND

FIG. 17

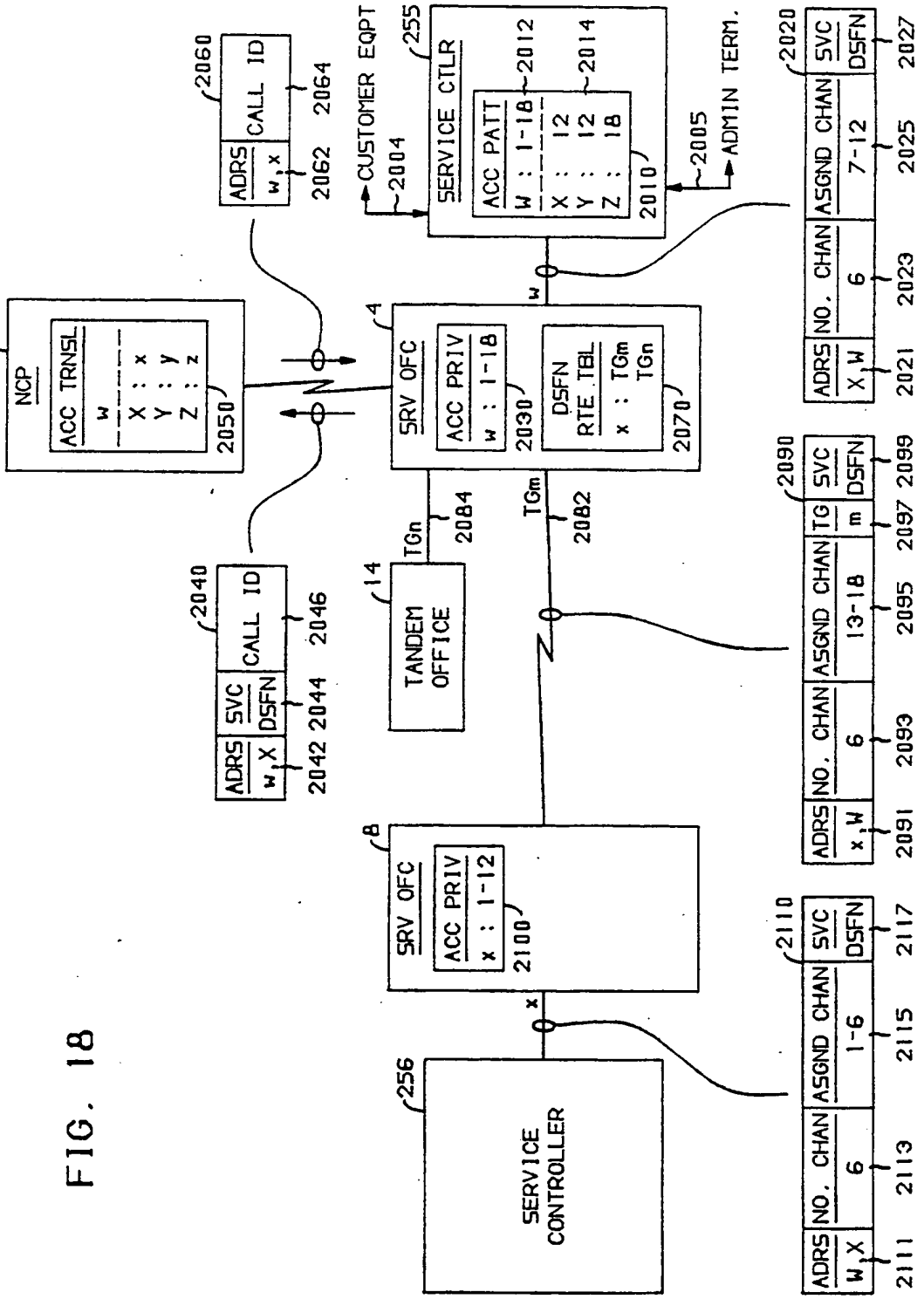
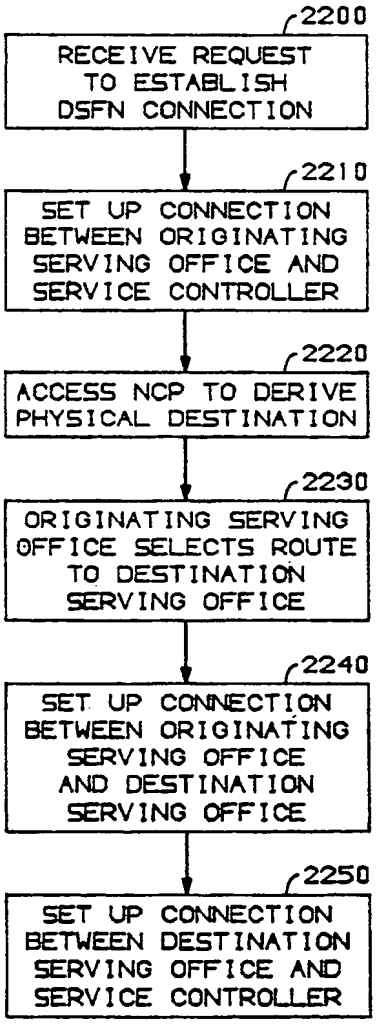


FIG. 18

FIG. 19


2200
RECEIVE REQUEST TO ESTABLISH DSFN CONNECTION
2210
SET UP CONNECTION BETWEEN ORIGINATING SERVING OFFICE AND SERVICE CONTROLLER
2220
ACCESS NCP TO DERIVE PHYSICAL DESTINATION
2230
ORIGINATING SERVING OFFICE SELECTS ROUTE TO DESTINATION SERVING OFFICE
2240
SET UP CONNECTION BETWEEN ORIGINATING SERVING OFFICE AND DESTINATION SERVING OFFICE
2250
SET UP CONNECTION BETWEEN DESTINATION SERVING OFFICE AND SERVICE CONTROLLER

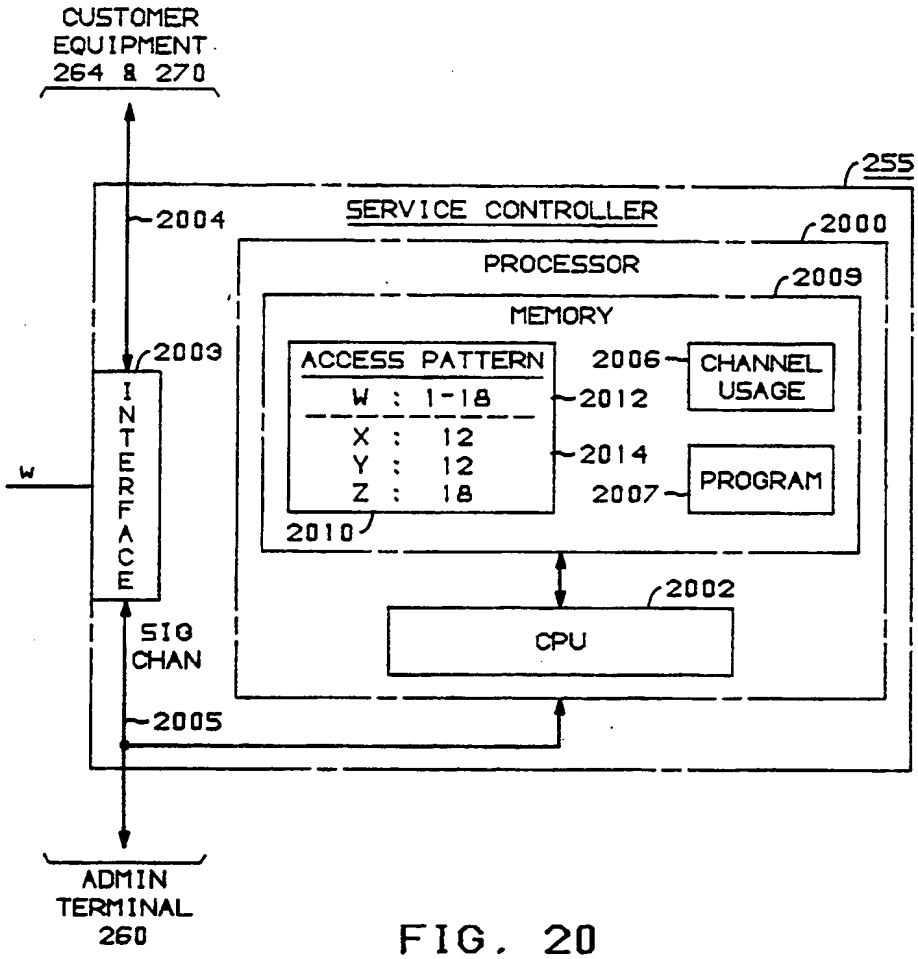


FIG. 20

This Page Blank (USPTO)

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 400 879 A3**

(12) EUROPEAN PATENT APPLICATION

(21) Application number: **90305585.3**

(51) Int. Cl.5: **H04Q 3/66**, H04Q 3/00

(22) Date of filing: **23.05.90**

(30) Priority: **30.05.89 US 359015**

(43) Date of publication of application:
05.12.90 Bulletin 90/49

(84) Designated Contracting States:
DE FR GB IT

(88) Date of deferred publication of the search report:
16.12.92 Bulletin 92/51

(71) Applicant: **AMERICAN TELEPHONE AND TELEGRAPH COMPANY**
550 Madison Avenue
New York, NY 10022(US)

(72) Inventor: **Gordon, Travis Hill**
41 Winding Way
Madison, New Jersey 07940(US)

(74) Representative: **Buckley, Christopher Simon Thirsk et al**
AT&T (UK) LTD. AT&T Intellectual Property Division 5 Mornington Road
Woodford Green, Essex IG8 OTU(GB)

(54) Dynamic shared facility system for private networks.

(57) This invention relates to a dynamically shared facility network (DSFN) providing private network service to a plurality of customers using switched facilities of a common carrier network. A plurality of serving offices (2,4,6,8,10,12',64) are connected via access links to customer telecommunications equipment (1). A pool of channels is dedicated to providing communications for private network service among these serving offices. In response to a request from a customer, connections are set up in the serving offices between access links and members of the pool of channels, in order to interconnect the serving links sought to be connected by the request. Where tandem connections between serving offices are necessary, connections are set up between members of the pool of channels. In case of failure of one or more channels, a new connection is automatically established. Advantageously, communication channels of the large communications facilities of a public switched network can be allocated to the DSFN, thus achieving economies of scale, and thus permitting use of the large and flexible switching systems of the public switched network to control and switch channels of the DSFN.

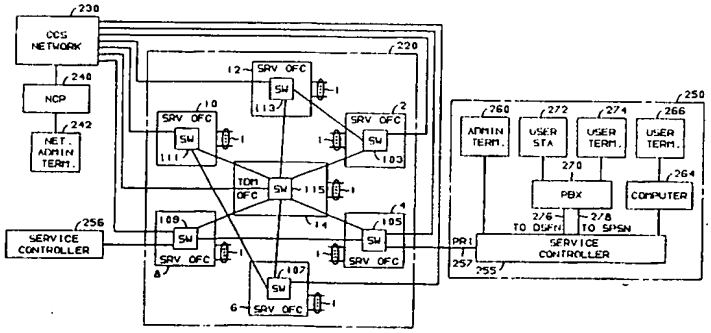


FIG. 2

Rank Xerox (UK) Business Services

European Patent Office

EUROPEAN SEARCH REPORT

Application Number

EP 90 30 5585

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl.5)
X	US-A-4 348 554 (ASMUTH) * Whole document *	1,3,5-11,14,15,17,18,20-27,32,34	H 04 Q 3/66 H 04 Q 3/00
A		36	
A	TENCON'87, Seoul, 25th - 28th August 1987, session 16, paper 1, vol. 2, pages 1-6; B. ERICSON: "Requirements on switching systems to be used in modern networks" * Page 3, left-hand column; pages 4-5; figures *	1,9,10,32,36	
A	NATIONAL TELECOMMUNICATIONS CONFERENCE, New Orleans, Louisiana, 29th November - 3rd December 1981, session G7, paper 6, vol. 4, pages 1-5; P.T. DE SOUSA: "Selection of on-net locations in private networks" * Whole document *	1,9,10,32,36	TECHNICAL FIELDS SEARCHED (Int. Cl.5) H 04 Q
A	REVIEW OF THE ELECTRICAL COMMUNICATION LABORATORIES, vol. 36, no. 1, January 1988, pages 41-48, Tokyo, JP; I. TOKIZAWA et al.: "An advanced multimedia TDM system for closed nerworks" * Whole document *	1,3,5-8,11,14,15,17,18,20-27,32,34,36	

~~The present search report has been drawn up for all claims~~

Place of search	Date of completion of the search	Examiner
THE HAGUE	27-05-1992	KURVERS F.J.J.

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P0401)

European Patent Office

CLAIMS INCURRING FEES

The present European patent application comprised at the time of filing more than ten claims.

☐ All claims fees have been paid within the prescribed time limit. The present European search report has been drawn up for all claims.

☐ Only part of the claims fees have been paid within the prescribed time limit. The present European search report has been drawn up for the first ten claims and for those claims for which claims fees have been paid,

namely claims:

☐ No claims fees have been paid within the prescribed time limit. The present European search report has been drawn up for the first ten claims.

LACK OF UNITY OF INVENTION

The Search Division considers that the present European patent application does not comply with the requirement of unity of invention and relates to several inventions or groups of inventions,
namely:

see sheet -B-

☐ All further search fees have been paid within the fixed time limit. The present European search report has been drawn up for all claims.

☐ Only part of the further search fees have been paid within the fixed time limit. The present European search report has been drawn up for those parts of the European patent application which relate to the inventions in respect of which search fees have been paid,

namely claims:

☒ None of the further search fees has been paid within the fixed time limit. The present European search report has been drawn up for those parts of the European patent application which relate to the invention first mentioned in the claims,

namely claims: point 1.

LACK OF UNITY OF INVENTION

The Search Division considers that the present European patent application does not comply with the requirement of unity of invention and relates to several inventions or groups of inventions, namely:

1. Claims: 1,3,5-11,14,15,17,18,20-27,32,34,36 as far as it concerns this subject-matter
   Sharing network facilities amongst several private network (and public network) customers.

2. Claims: 2,12,13,28-31,33,36 as far as it concerns this subject matter
   Failure rerouting/rearranging of virtual private facilities

3. Claims: 4,16,19,35,36 as far as it concerns this subject-matter
   Signalling channels in a (virtual) network

98/1074

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 696 147 A1

(12) EUROPÄISCHE PATENTANMELDUNG

(43) Veröffentlichungstag:
07.02.1996 Patentblatt 1996/06

(51) Int. Cl.[6]: **H04Q 3/66**

(21) Anmeldenummer: **94112147.7**

(22) Anmeldetag: **03.08.1994**

(84) Benannte Vertragsstaaten:
AT BE CH DE DK ES FR GB GR IE IT LI LU NL PT SE

(71) Anmelder: **SIEMENS AKTIENGESELLSCHAFT**
D-80333 München (DE)

(72) Erfinder:
- **Stademann, Rainer, Dr. rer. nat.**
  D-85658 Egmating (DE)
- **Gehlhaus, Karl, Ing.**
  D-81377 München (DE)

(54) Verfahren und Routing-System zur dynamischen Verkehrslenkung in einem Kommunikationsnetz

(57) Die dynamische Leitweglenkung eines Kommunikationsnetzes soll sich an die jeweilige im Netz vorliegende Verkehrslast so anpassen, daß der Netzdurchsatz optimiert wird.

Die erfindungsgemäße Leitweglenkung löst diese Problem durch einen Wegefächer zur Aufnahme von Alternativwegen für den Überlaufverkehr, wobei ein bisher im Wegefächer enthaltenerr Alternativweg ersatzlos aus dem Wegefächer entfernt wird, sobald festgestellt wird, daß er nicht mehr verfügbar ist.

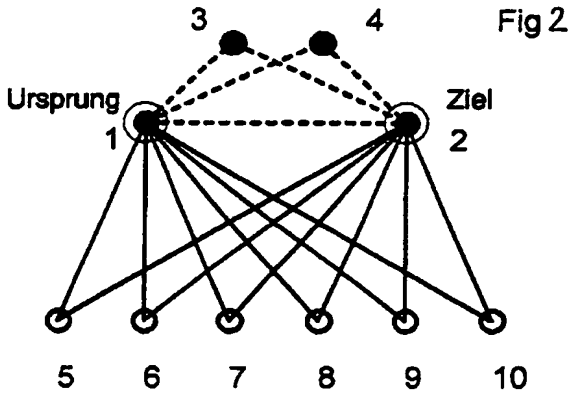


Fig 2

EP 0 696 147 A1

Beschreibung

Die Erfindung betrifft ein Verfahren und ein Routing-System zur dynamischen Verkehrslenkung in einem Kommunikationsnetz

Nichthierarchisch organisierte, leitungsvermittelnde Kommunikationsnetze benötigen eine dynamische Leitweglenkung (dynamic routing), die sich an die jeweilige im Netz vorliegende Verkehrslast so anpaßt, daß der Netzdurchsatz optimiert wird. Dabei müssen insbesondere auch Schieflastsituationen durch die Leitweglenkung entschärft werden.

So wie bei der konventionellen Leitweglenkung in hierarchischen Netzen wird auch bei einer dynamischen Leitweglenkung erst versucht, Verbindungen über einen oder mehrere Planwege, die zumeist Direktwege sind, aufzubauen. Ist dies nicht möglich, weil z.B. alle Verbindungsleitungsbündel des Direktweges vollständig belegt sind, wird der Überlaufverkehr Alternativwegen zugewiesen.

In der Auswahl der Alternativwege liegt der prinzipielle Unterschied zwischen dynamischer und konventioneller Leitweglenkung. Bei der konventionellen Leitweglenkung werden administrativ festgelegte Alternativwege in starrer Reihenfolge nach einer unbelegten Leitung bzw. einem unbelegten Kanal abgesucht ("fixed alternate routing"). Dadurch kann die Leitweglenkung nur sehr ungenügend auf nicht geplante, außergewöhnliche Lastsituationen reagieren.

Bei der dynamischen Leitweglenkung wird entstehender Überlaufverkehr einem oder mehreren aktiven Alternativwegen zugewiesen. Dieser aktive Alternativweg bzw. diese aktiven Alternativwege sind nicht fest, sondern werden entsprechend dem jeweiligen Verfahren zur dynamischen Leitweglenkung ausgewählt oder sogar bei jedem Call neu bestimmt. Die Vorteile der dynamischen Leitweglenkung liegen in Robustheit und Flexibilität gegenüber Schieflastsituationen im Netz, die z.B. durch zeitlich schwankende Lasten (räumlich begrenztes starkes Verkehrsaufkommen z.B. bei Katastrophen) und Netzdegradierung (Verbindungsleitungsbündelausfälle, Ausfälle von Vermittlungseinheiten) entstehen können. Außerdem können Unsicherheiten bei der Netzplanung besser kompensiert werden.

Aus der europäischen Patentschrift EP - B1 0 229 494 ist ein dezentrales Verfahren zur dynamischen Leitweglenkung bekannt, das einem Alternativweg solange Überlaufverkehr zuweist, bis dieser nicht mehr verfügbar ist, d.h. bis entweder der Erst-Link des Alternativwegs belegt ist, oder der Ursprungsknoten eine Auslösungsmeldung wegen Blockierung von einem Transitknoten erhält. In diesem Fall wird der bisherige Alternativweg zyklisch oder (pseudo-) zufällig durch einen anderen Alternativweg ersetzt. In einer anderen Ausprägung des Verfahrens wird der Überlaufverkehr auf eine Gruppe von mehreren Alternativwegen verteilt und ein Alternativweg im Falle eines Nichtverfügbarwerdens durch einen anderen Alternativweg ersetzt.

Das genannte Verfahren hat den Nachteil, daß auch hochbelastete Alternativwege immer wieder Überlaufverkehr erhalten, selbst wenn noch niedrig belastete Alternativwege zur Verfügung stehen würden und der entstehende Überlaufverkehr nicht allen wenig belasteten Wegen gleichmäßig angeboten wird.

Der Erfindung liegt die Aufgabe zugrunde, den Überlaufverkehr gleichmäßig auf möglichst niedrigbelastete Alternativwege zu verteilen.

Diese Aufgabe wird durch die Merkmale des Anspruchs 1 bzw. des Anspruchs 9 gelöst.

Durch das Nichtersetzen eines wegen Nichtverfügbarkeit aus dem Wegefächer herausgenommenen Alternativweges wird verhindert, daß hochbelastete Alternativwege zu früh nach ihrem Ausscheiden bereits wieder für den Überlaufverkehr angeboten werden und damit wieder Verkehr erhalten, obwohl noch niedrig belastete Alternativwege zur Verfügung stehen würden.

Des weiteren ist das erfindungsgemäße Verfahren weniger zeitaufwendig, da nach einem Feststellen der Nichtverfügbarkeit eines Alternativweges nicht jedes Mal ein Ersatz-Alternativweg bestimmt werden muß.

In einer Ausgestaltung der Erfindung nach Anspruch 3 besteht der Wegefächer der aktiven Alternativwege nach jeder (Re-) Initialisierung aus allen für die Ursprungs-Ziel-Beziehung möglichen Alternativwegen. Die (Re-)Initialisierung ist dadurch sehr wenig aufwendig.

In einer weiteren Ausgestaltung der Erfindung nach Anspruch 4 besteht der Wegefächer der aktiven Alternativwege nach jeder (Re-)Initialisierung aus einer echten Untermenge der für die Ursprungs-Ziel-Beziehung möglichen Alternativwege. Alternativwege, von denen a priori bekannt ist, daß sie momentan oder ständig wenig freie Kapazität besitzen, können damit bereits bei der Initialisierung des Verfahrens aus dem Wegefächer ausgeschlossen werden.

Im folgenden wird ein Ausführungsbeispiel des erfindungsgemäßen Verfahrens anhand FIG 1 näher erläutert.

FIG 1 zeigt ein Diagramm eines kleinen vollvermaschten Netzwerks mit fünf Netz-Vermittlungsknoten und den entsprechenden Kapazitäten der Wegabschnitte(Links) zwischen den Netzknoten, wobei ein Link mindestens ein Verbindungsleitungsbündel umfaßt.

Es werde nun angenommen, daß der Vermittlungsknoten 1 einen Call für Vermittlungsknoten 2 hat, aber die direkte Route zwischen den beiden Vermittlungsknoten nicht verfügbar ist.

Weiter werde angenommen, daß der initiale Wegefächer, d.h. der Fächer der aktiven Alternativwege nach seiner erstmaligen Initialisierung oder nach einer Reinitialisierung, aus allen für die Ursprungs-Ziel-Beziehung möglichen Zweilink-Alternativwegen besteht. Unter dieser Voraussetzung umfaßt der Wegefächer der Alternativwege für die Ursprungs-Ziel-Beziehung zwischen Vermittlungsknoten 1 und Vermittlungsknoten 2 drei Alternativwege, nämlich die Zweilink-Alternativwege über die Vermittlungsknoten 3, 4 und 5.

Es werde weiterhin davon ausgegangen, daß der überlaufende Verkehr vom Routing-System zyklisch auf diese aktiven Alternativwege gleichmäßig verteilt wird und zwar in der Reihenfolge Transitknoten 3, 4 und 5.

Unter den genannten Voraussetzungen überprüft das Routing-System am Vermittlungsknoten 1 zunächst, ob der aktive Alternativweg über Transitknoten 3 verfügbar ist, d.h. ob er belegbare freie Leitungen bzw. Kanäle aufweist (im folgenden wird nur von "Kanälen" gesprochen).

Um dies überprüfen zu können, speichert das Routing-System in Vermittlungsknoten 1 die Kapazität des Links zwischen Vermittlungsknoten 1 und 3, nämlich 125 Kanäle und den für diesen Link zugehörigen Trunk-Reservation-Parameter, der hier beispielsweise 10 sei. Das Routing-System speichert darüber hinaus die Anzahl der momentan benutzten Kanäle. Der Link zwischen Vermittlungsknoten 1 und 3 ist aus der Sicht des Routing-System dann für Überlaufverkehr verfügbar, wenn die Summe der benutzten Kanäle und des Trunk-Reservation-Parameters kleiner als 125 ist (Die Trunkreservierung garantiert die Stabilität eines Routing-Verfahrens im Hochlastbereich).

Ist der erste Link verfügbar, so baut der Vermittlungsknoten 1 die Verbindung zunächst bis zum Vermittlungsknoten 3 auf. Das Routing-System des Vermittlungsknotens 3 prüft dann vor der Weiterführung des Verbindungsaufbaus zum Ziel-Vermittlungsknoten 2 die Verfügbarkeit des zweiten Links, indem es überprüft, ob die Summe von belegte Kanäle plus Trunk-Reservation-Parameter kleiner als die Kapazität des zweiten Links ist (Das Routing-System des Vermittlungsknotens 3 kennt hierzu die Kapazität des Links zwischen Vermittlungsknoten 3 und 2, nämlich 125 Kanäle, sowie den Trunk-Reservation Parameter dieses Links, nämlich 10 Kanäle und die Anzahl der momentan belegten Kanäle dieses Links).

Falls auch der zweite Link des genannten aktiven Alternativweges verfügbar ist, wird die Verbindung vom Transitknoten 3 zum Zielknoten 2 aufgebaut.

Falls der Transitknotgen 3 feststellt, daß der Link zum Zielknoten 2 nicht verfügbar ist, lost Transitknoten 3 den Verbindungsabschnitt zum Ursprungsknoten 1 mit einer speziell gekennzeichneten Rückwärtsmeldung (crankback-Meldung) aus. Das Routing-System des Ursprungsknotens 1 entfernt daraufhin den Alternativweg über Transitknoten 3 aus dem Wegefächer für Zielknoten 2.

Im vorhergenannten Fall der erfolgreichen Vermittlung des Calls über Transitknoten 3 wird beim nächsten Call für den Vermittlungsknoten 2 bei Nichtverfügbarkeit der direkten Route nochmals versucht, den Call über den Transitknoten 3 zu lenken. Erst bei einem weiteren Call wird dann zyklisch gewechselt, d.h. der Call wird über den nächsten aktiven Alternativweg gelenkt, d.h. den aktiven Alternativweg über Transitknoten 4. Dadurch können kurzzeitige Autokorrelationen im Verkehrsangebot auf dem zweiten Link genutzt werden, die die Wahrscheinlichkeit erhöhen, daß unmittelbar nach einem erfolgreichen Verbindungsaufbau eine weitere Verbindung auf dem gleichen Weg aufgebaut werden kann.

Bei dem Ausführungsbeispiel wird der vom Direktweg überlaufende Verkehr also zyklisch umlaufend den aktiven Alternativwegen zugewiesen. Dabei erhält jeder der aktiven Alternativwege zwei aufeinander folgende, vom Direktweg überlaufende Calls zugewiesen.

Sobald das-Routing-System bei der Durchführung des Routing-Verfahrens feststellt, daß ein aktiver Alternativweg des Wegefächers nicht mehr verfügbar ist, wird dieser aus dem Wegefächer entfernt, jedoch nicht durch einen anderen Alternativweg ersetzt (unter dem Entfernen des Alternativwegs aus dem Wegefächer kann auch verstanden werden, daß der betroffene Alternativweg durch ein Kennzeichen als nicht verfügbar markiert wird). Durch das Nichtersetzen des entfernten Alternativweges im Wegefächer wird vermieden, daß ein Alternativweg mit unter Umständen wenig freien Leitungen durch Ersetzen eines anderen nicht mehr verfügbar gewordenen Alternativwegs wieder in die Wegesequenz aufgenommen wird und somit Verkehr auf dem neu aufgenommenen Alternativweg verloren geht.

Stellt das Routing-System fest, daß es einen Call einem Alternativweg mit einem besetzten ersten Wegabschnitt angeboten (zugewiesen) hat, wird der aktive Alternativweg gewechselt und der Call wird dem zyklisch nächsten aktiven Alternativweg angeboten. Ist auch hier der erste Wegabschnitt besetzt, wird der Alternativweg ein weiteres Mal getauscht. Insgesamt wird für einen Call maximal eine vorgegebene Anzahl von aktiven Alternativwegen auf verfügbare Kanäle im ersten Wegabschnitt geprüft, bevor der Call zu Verlust geht. Ein Call, der bei verfügbarem ersten Wegeabschnitt auf einen nicht verfügbaren zweiten Wegeabschnitt trifft, geht bei dem Ausführungsbeispiel sofort zu Verlust (kein "Rerouting"). Die Erfindung kann jedoch auch mit "Rerouting" realisiert werden.

Alternativwege mit wenig freien Kanälen werden durchschnittlich schneller nicht verfügbar als Alternativwege mit viel freien Kanälen. Letztere bleiben also durchschnittlich länger im Wegefächer der aktiven Alternativwege und erhalten somit auch mehr Überlaufverkehr zugewiesen.

Da der Überlaufverkehr auf alle im Wegefächer verbliebenen aktiven Alternativwege verteilt wird, bleiben die Alternativwege länger verfügbar als bei einem Verfahren, das mit einer festen Anzahl von aktiven Alternativwegen arbeitet.

Erst wenn überhaupt kein aktiver Alternativweg mehr existiert oder eine vorgegebene Anzahl von aktiven Alternativwegen im Wegefächer unterschritten wird, reinitialisiert das Routing-System den genannten Wegefächer.

Durch das genannte Entfernen der nicht mehr verfügbaren aktiven Alternativwege aus dem Wegefächer werden die "schlechten" Alternativwege ausgesiebt. Dadurch gehen beim Verteilen des aufkommenden Überlaufverkehrs auf alle im Wegefächer verbliebenen

Alternativwege insbesondere unter Schieflast-Situationen im Netz nur sehr wenige Calls verloren.

In der Praxis verändert sich die Lastsituation im Netz ständig, -so daß ein Routing-System mit einem Wegefächer fester Größe nur selten eine optimale Verteilung des Überlaufverkehrs besitzen wird. Dagegen stellt das erfindungsgemäße Routing-System die Große des Wegefächers schnell auf einen optimalen Wert ein und hält diesen relativ lange Zeit.

Diese Wirkungsweise wird im folgenden anhand von FIG 2 näher erläutert.

FIG 2 zeigt einen Ausschnitt aus einem vollvermaschten Kommunikationsnetz mit zehn Vermittlungsknoten, wobei in dem Ausschnitt der Ein-Link-Direktweg für Verkehr vom Ursprungsknoten 1 zum Zielknoten 2 dargestellt ist und alle Zwei-Link-Alternativwege dieser Ursprungs-Ziel-Beziehung über die Transitknoten 3 bis 10.

Für das Beispiel in FIG 2 wird ebenfalls wie bei FIG 1 davon ausgegangen, daß als Alternativwege nur Zwei-Link-Wege in Frage kommen.

Im in FIG 2 dargestellten Beispiel sei angenommen, daß zwischen den Vermittlungsknoten 1 bis 4 (Netzbereich A) für einige Zeit ein außerplanmäßiges, stark erhöhtes Verkehrsangebot herrsche, während zwischen den Vermittlungsknoten 5 bis 10 (Netzbereich B) und zwischen den Vermittlungsknoten von Bereich A und Bereich B ein normales, planmäßiges Verkehrsangebot vorliegt. Durch das starke Verkehrsangebot innerhalb von Bereich A werden die gestrichelt eingezeichneten Links (aufgrund der Trunk-Reservierung) fast ausschließlich mit Direktwegverkehr belegt. Dadurch weisen die Alternativwege über Transitknoten 3 und 4 eine hohe Blockierungswahrscheinlichkeit für den Überlaufverkehr (z.B. 99 %) auf, während die sechs Alternativwege, die über die Transitknoten 5 bis 10 führen (in FIG 2 durchgezogen eingezeichnet) insgesamt eine sehr geringe Blockierungswahrscheinlichkeit (z.B. 0,01 %) haben.

Es wird nun wieder davon ausgegangen, daß der Wegefächer für den Verkehr von Vermittlungsknoten 1 zum Vermittlungsknoten 2 des dargestellten Beispiels mit dem vollen Wegefächer initialisiert wird, d.h. mit den in FIG 2 dargestellten acht Alternativwegen. Diesen Alternativwegen wird der vom Direktweg (1-2) überlaufende Verkehr bzw. die überlaufenden Calls angeboten. Da die Blockierungswahrscheinlichkeit auf den Alternativwegen über Transitknoten 3 und 4 wesentlich hoher ist, als auf den Alternativwegen über Transitknoten 5 bis 10, werden die beiden hochbelasteten Alternativwege bald aus dem Wegefächer entfernt. (Bei den oben angenommenen Blockierungswahrscheinlichkeiten wird ein hochbelasteter Alternativweg in 99 von 100 Fällen nach dem ersten zugeteilten Call entfernt, während dies bei einem der sechs niedrig belasteten Alternativwege nur in einem von 10 000 Fällen passiert).

Nachdem sich der Wegefächer gemäß der genannten Wirkungsweise auf die niedrig belasteten Alternativwege reduziert hat, wird der aufkommende Überlaufverkehr nur noch auf die sechs niedrig belasteten Alternativwege nach einem bestimmten Auswahlschema (zufallsgesteuert oder pseudozufallsgesteuert oder zyklisch umlaufend) gleichmäßig verteilt. Der Anstieg der Blockierungswahrscheinlichkeit durch den zugeteilten Überlaufverkehr auf den Alternativwegen wird also minimiert.

Der Wegefächer reduziert sich also sehr schnell auf die optimale Große und verkleinert sich dann nur noch langsam.

Verändert sich die im Beispiel betrachtete Netzlast dahingehend, daß ein bisher niedrig belasteter Alternativweg hoch belastet wird, wird dieser bald aus dem Wegefächer entfernt. Wird umgekehrt ein bisher hochbelasteter Alternativweg zu einem niedrig belasteten Alternativweg, stellt sich nach der nächsten Reinitialisierung der Wegefächer schnell wieder auf die neue optimale Größe ein.

Im Gegensatz dazu kann sich bei Verfahren, die mit Wegefächern fester Größe arbeiten, die Größe der Fächer nicht an die sich verändernden Verkehrsschieflasten im Netz anpassen. Dadurch enthält ein Wegefächer fester Größe oft entweder hochbelastete Alternativwege oder nicht alle niedrig belasteten Alternativwege. Im ersten Fall geht Verkehr auf den hochbelasteten Alternativwegen verloren. Im zweiten Fall wird der aufkommende Überlaufverkehr ungleichmäßig auf die niedrig belasteten Wege verteilt, so daß wiederum mehr Verkehr als beim erfindungsgemäßen Verfahren verlorengeht.

Wird der Wegefächer einer Ursprungs-Ziel-Beziehung nur dann reinitialisiert, wenn die Zahl der in ihm enthaltenen Alternativwege eine vorgegebene Anzahl unterschreitet, so kann die Zeitkonstante des durch das Nichtersetzen eines aus dem Wegefächer entfernten Alternativweges erzielten Siebvorgangs bei geringem Überlaufverkehr recht lang werden. Durch eine zusätzlich rein zeitlich bedingte Reinitialisierung (Ablauf einer bestimmten Zeitspanne, z.B. 10-15 Min., seit der letzten Realisierung oder periodische Reinitialisierung, angestoßen durch ein netzzentralen Verkehrsmanagement-System unter Berücksichtigung der aktuellen Lastsituation im Netz) kann die Möglichkeit für eine derartig lange Zeitkonstante vermieden und somit erreicht werden, daß Alternativwege, die aus dem Wegefächer gefallen sind, aber inzwischen durch Veränderung der Netzlast wieder freie Kapazität besitzen, früher wieder in den Wegefächer aufgenommen werden.

Des weiteren ist es auch möglich, daß das genannte netzzentrale Verkehrsmanagement-System nach dem Erkennen einer Schieflastsituation im Netz eine Reinitialisierung der Wegefächer aperiodisch anstößt.

Schließlich kann auch der Netzbetreiber eine aperiodische Reinitialisierung anstoßen.

Patentansprüche

1. Verfahren zur dynamischen Verkehrslenkung in einem Kommunikationsnetz, demgemäß

a) Calls zwischen einem Ursprungs-Vermittlungsknoten und einem Ziel-Vermittlungsknoten zunächst einem oder mehreren bevorzugten Wegen (Planwegen) angeboten werden,
b) für den Fall, daß keiner der Planwege verfügbar ist, Calls Alternativwegen, die in einem Wegefächer enthalten sind, nach einem bestimmten Auswahlschema angeboten werden,
c) ein bisher im Wegefächer enthaltener Alternativweg aus dem Wegefächer entfernt wird, sobald festgestellt wird, daß er nicht mehr verfügbar ist,
d) der aus dem Wegefächer entfernte Alternativweg nicht ersetzt wird,
e) der Wegefächer periodisch, d.h. nach bestimmten Zeitabständen, und/oder aperiodisch, d.h. nach Eintreten wenigstens eines Ereignisses oder aufgrund eines Kommandos von einem netzzentralen Verkehrsmanagement-System oder dem Netzbetreiber reinitialisiert wird.

2. Verfahren nach Anspruch 1,
dadurch gekennzeichnet,
daß als Ereignis nach dessen Eintreten reinitialisiert wird das Unterschreiten der in dem Wegefächer enthaltenen Zahl von Alternativwegen unter eine bestimmte Anzahl oder das Ablaufen einer bestimmten Zeitspanne seit der letzten Reinitialisierung in Frage kommt.

3. Verfahren nach Anspruch 1 oder 2,
dadurch gekennzeichnet,
daß dem Wegefächer bei einer Initialisierung bzw. Reinitialisierung alle möglichen Alternativwege zugeordnet werden.

4. Verfahren nach Anspruch 1 oder 2,
dadurch gekennzeichnet,
daß dem Wegefächer bei einer Initialisierung bzw. Reinitialisierung eine Untermenge der möglichen Alternativwege zugeordnet wird, die nach dem Kriterium der Verkehrsbelastung oder der freien Kapazität der Alternativwege von dem Ursprungs-Vermittlungsknoten oder einem netzzentralen Verkehrsmanagement-System ermittelt wird.

5. Verfahren nach einem der Ansprüche 1 bis 4,
dadurch gekennzeichnet,
daß ein Call nur einem einzigen Alternativweg angeboten wird, bevor er wegen dessen Nichtverfügbarkeit ausgelöst wird.

6. Verfahren nach einem der Ansprüche 1 bis 4,
dadurch gekennzeichnet,
daß ein Call mehreren Alternativwegen angeboten wird, bevor er wegen deren Nichtverfügbarkeit ausgelöst wird.

7. Verfahren nach einem der Ansprüche 1 bis 6,
dadurch gekennzeichnet,
daß das genannte bestimmte Auswahlschema darin besteht, daß die Alternativwege aus dem Wegefächer zufallsgesteuert oder pseudozufallsgesteuert oder zyklisch umlaufend ausgewählt werden.

8. Verfahren nach einem der Ansprüche 1 bis 7,
dadurch gekennzeichnet,
daß jedem aus dem Wegefächer für den Überlaufverkehr ausgewählten Alternativweg eine vorgegebene Anzahl von überlaufenden Calls angeboten wird, bevor zu dem nach dem Auswahlschema nächsten Alternativweg übergegangen wird.

9. Routing-System zur dynamischen Verkehrslenkung in einem Vermittlungsknotenprozessor eines Kommunikationsnetzes, das derart ausgestaltet ist, daß es

a) Calls zwischen einem Ursprungs-Vermittlungsknoten und einem Ziel-Vermittlungsknoten zunächst einem oder mehreren bevorzugten Wegen(Planwegen) anbietet,
b) für den Fall, daß keiner der Planwege verfügbar ist, Calls Alternativwegen, die in einem Wegefächer enthalten sind, nach einem bestimmten Auswahlschema anbietet,
c) einen bisher im Wegefächer enthaltenen Alternativweg aus dem genannten Wegefächer entfernt, sobald es feststellt, daß er nicht mehr verfügbar ist,
d) den aus dem Wegefächer entfernten Alternativweg nicht ersetzt,
e) den Wegefächer periodisch, d.h. nach bestimmten Zeitabständen, und/oder aperiodisch, d.h. nach Eintreten wenigstens eines Ereignisses oder aufgrund eines Kommandos von einem netzzentralen Verkehrsmanagement-System oder dem Netzbetreiber reinitialisiert.

10. Routing-System nach Anspruch 9,
dadurch gekennzeichnet,
daß als Ereignis nach dessen Eintreten das Routing-System den Wegefächer reinitialisiert das Unterschreiten der in dem Wegefächer enthaltenen Zahl von Alternativwegen unter eine bestimmte Anzahl oder das Ablaufen einer bestimmten Zeitspanne seit der letzten Reinitialisierung in Frage kommt.

11. Routing-System nach Anspruch 9 oder 10,
dadurch gekennzeichnet,
daß es dem Wegefächer bei einer Initialisierung

bzw. Reinitialisierung alle möglichen Alternativwege zuordnet.

12. Routing-System nach Anspruch 9 oder 10,
dadurch gekennzeichnet,
daß es dem Wegefächer bei einer Initialisierung bzw Reinitialisierung eine Untermenge-der möglichen Alternativwege zuordnet, die nach dem Kriterium der Verkehrsbelastung oder der freien Kapazität der Alternativwege von dem Routing-System selbst oder einem netzzentralen Verkehrsmanagement-System ermittelt wird.

13. Routing-System nach einem der Ansprüche 9 bis 12,
dadurch gekennzeichnet,
daß es einen Call einem einzigen Alternativweg anbietet, bevor er bei dessen Nichtverfügbarkeit ausgelost wird.

14. Routing-System nach einem der Ansprüche 9 bis 12,
dadurch gekennzeichnet,
daß es einen Call mehreren Alternativwegen anbietet, bevor er bei deren Nichtverfügbarkeit ausgelost wird.

15. Routing-System nach einem der Ansprüche 9 bis 14,
dadurch gekennzeichnet,
das genannte bestimmte Auswahlschema darin besteht, daß die Alternativwege aus dem Wegefächer zufallsgesteuert oder pseudozufallsgesteuert oder zyklisch umlaufend ausgewählt werden.

16. Routing-System nach einem der Ansprüche 9 bis 15,
dadurch gekennzeichnet,
daß es jedem aus dem Wegefächer für den Überlaufverkehr ausgewählten Alternativweg eine vorgegebene Anzahl von überlaufenden Calls anbietet, bevor es den nach dem Auswahlschema nächsten Alternativweg anbietet.

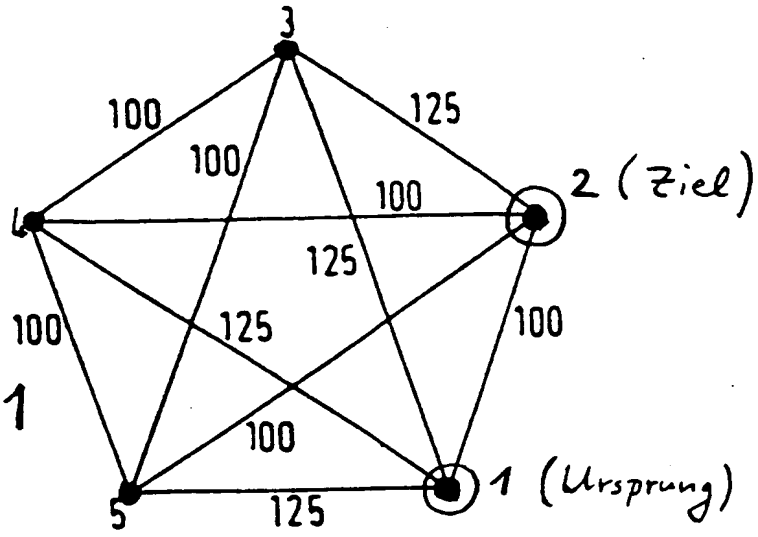


FIG.1

Link capacities

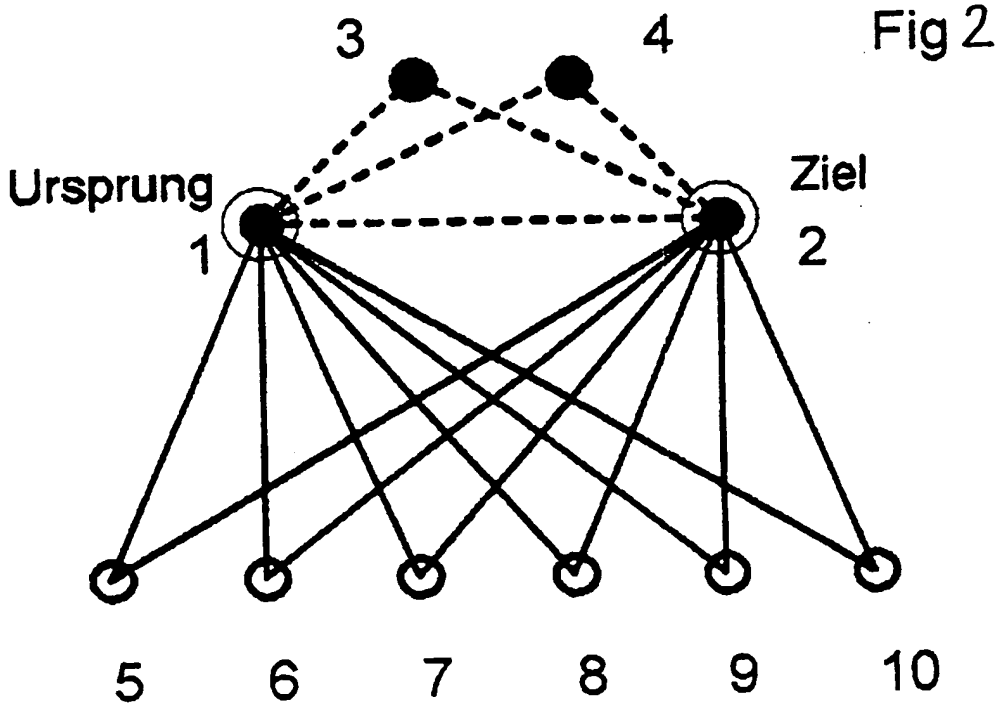


Fig 2

● Vermittlungsknoten Netzbereich A

○ Vermittlungsknoten Netzbereich B

Europäisches Patentamt

EUROPÄISCHER RECHERCHENBERICHT

Nummer der Anmeldung
EP 94 11 2147

EINSCHLÄGIGE DOKUMENTE

Kategorie	Kennzeichnung des Dokuments mit Angabe, soweit erforderlich, der maßgeblichen Teile	Betrifft Anspruch	KLASSIFIKATION DER ANMELDUNG (Int.Cl.6)
X	PROCEEDINGS, 10TH INTERNATIONAL TELETRAFFIC CONGRESS, 9-15 JUNI 1983, HEFT 1 SITZUNG 3.2 DOKUMENT 3 SEITEN 1-8, MONTREAL CN W.H. CAMERON ET AL 'Dynamic Routing for Intercity Telephone Networks' * Zusammenfassung * * Seite 1, rechte Spalte, Zeile 25 - Zeile 42 * * Seite 2, rechte Spalte, Zeile 8 - Zeile 21 * * Seite 3, linke Spalte, Zeile 24 - Zeile 29 * * Seite 8, linke Spalte, Zeile 6 - Zeile 15 * ---	1-16	H04Q3/66
X	EP-A-0 372 270 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) * Zusammenfassung * * Seite 3, Zeile 53 - Seite 4, Zeile 5 * * Seite 7, Zeile 39 - Zeile 46; Abbildung 1 * ---	1-16	RECHERCHIERTE SACHGEBIETE (Int.Cl.6) H04Q
A	EP-A-0 490 446 (KONINKLIJKE PTT NEDERLAND N.V.) * Zusammenfassung * * Spalte 2, Zeile 16 - Zeile 48 * * Spalte 3, Zeile 55 - Spalte 4, Zeile 6; Abbildungen 1-4 * ---	1-16	
A	EP-A-0 376 556 (AMERICAN TELEPHONE AND TELEGRAPH COMPANY) * Zusammenfassung * * Seite 2, Zeile 49 - Seite 3, Zeile 35; Abbildung 1 * --- -/--	1-16	

Der vorliegende Recherchenbericht wurde für alle Patentansprüche erstellt

Recherchenort	Abschlußdatum der Recherche	Prüfer
DEN HAAG	23. Dezember 1994	O'Reilly, D

KATEGORIE DER GENANNTEN DOKUMENTE

X : von besonderer Bedeutung allein betrachtet
Y : von besonderer Bedeutung in Verbindung mit einer anderen Veröffentlichung derselben Kategorie
A : technologischer Hintergrund
O : nichtschriftliche Offenbarung
P : Zwischenliteratur

T : der Erfindung zugrunde liegende Theorien oder Grundsätze
E : älteres Patentdokument, das jedoch erst am oder nach dem Anmeldedatum veröffentlicht worden ist
D : in der Anmeldung angeführtes Dokument
L : aus andern Gründen angeführtes Dokument

& : Mitglied der gleichen Patentfamilie, übereinstimmendes Dokument

EPO FORM 1503 03.82 (P04C03)

Europäisches Patentamt

EUROPÄISCHER RECHERCHENBERICHT

Nummer der Anmeldung
EP 94 11 2147

EINSCHLÄGIGE DOKUMENTE

Kategorie	Kennzeichnung des Dokuments mit Angabe, soweit erforderlich, der maßgeblichen Teile	Betrifft Anspruch	KLASSIFIKATION DER ANMELDUNG (Int.Cl.6)
A	IEEE COMMUNICATIONS MAGAZINE, Bd.25, Nr.9, September 1987, PISCATAWAY, NJ US Seiten 13 - 21 B.R. HURLEY ET AL 'A Survey of Dynamic Routing Methods for Circuit-Switched Traffic' * Seite 17, linke Spalte, Zeile 4 - rechte Spalte, Zeile 11; Abbildung 4 *	1-16	
A	IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS, Bd.15, Nr.6, November 1985, NEW YORK US Seiten 730 - 736 B. AKSELROD ET AL 'A Simulation Study of Advanced Routing Methods in a Multipriority Telephone Network' * Zusammenfassung * * Seite 731, linke Spalte, Zeile 51 - Seite 732, Zeile 18 * * Seite 734, linke Spalte, Zeile 19 - Zeile 29; Abbildungen 1,3 *	1-16	RECHERCHIERTE SACHGEBIETE (Int.Cl.6)
A	IEEE COMMUNICATIONS MAGAZINE, Bd.28, Nr.10, Oktober 1990, PISCATAWAY, NJ US Seiten 54-58 - 63-64, XP165755 P.B. KEY ET AL 'Distributed Dynamic Routing Schemes' * das ganze Dokument *	1-16	

Der vorliegende Recherchenbericht wurde für alle Patentansprüche erstellt

Recherchenort	Abschlußdatum der Recherche	Prüfer
DEN HAAG	23. Dezember 1994	O'Reilly, D

KATEGORIE DER GENANNTEN DOKUMENTE

X : von besonderer Bedeutung allein betrachtet
Y : von besonderer Bedeutung in Verbindung mit einer anderen Veröffentlichung derselben Kategorie
A : technologischer Hintergrund
O : nichtschriftliche Offenbarung
P : Zwischenliteratur

T : der Erfindung zugrunde liegende Theorien oder Grundsätze
E : älteres Patentdokument, das jedoch erst am oder nach dem Anmeldedatum veröffentlicht worden ist
D : in der Anmeldung angeführtes Dokument
L : aus andern Gründen angeführtes Dokument

& : Mitglied der gleichen Patentfamilie, übereinstimmendes Dokument

EPO FORM 1503 03.82 (P04C03)

```
?b wpi

       06jul00 10:54:57 User212334 Session D2254.1
       Sub account: P001131
            $0.00    0.163 DialUnits FileHomeBase
     $0.00  Estimated cost FileHomeBase
     $0.08  TYMNET
     $0.08  Estimated cost this search
     $0.08  Estimated total session cost   0.163 DialUnits


File 351:DERWENT WPI  1963-2000/UD=, UM=, & UP=200030
       (c) 2000 Derwent Info Ltd
*File 351: Display format changes now online.
 Please see HELP NEWS 351 for details.

      Set  Items  Description
      ---  -----  -----------
?s pn=ep 696147
      S1       1  PN=EP 696147
?t s1/5

 1/5/1
DIALOG(R)File 351:DERWENT WPI
(c) 2000 Derwent Info Ltd. All rts. reserv.

010592220    **Image available**
WPI Acc No: 1996-089173/199610
XRPX Acc No: N96-074707
 Dynamic traffic routing system for communications network - has
 alternative call connection paths established when defined connection
 paths between source and target exchange are fully occupied
Patent Assignee: SIEMENS AG (SIEI  )
Inventor: GEHLHAUS K; STADEMANN R; GELHAUS K
Number of Countries: 023  Number of Patents: 008
Patent Family:
Patent No     Kind   Date     Applicat No     Kind   Date     Week
EP 696147      A1  19960207  EP 94112147      A   19940803  199610  B
WO 9604757     A1  19960215  WO 95EP3098      A   19950803  199613
FI 9700392     A   19970130  WO 95EP3098      A   19950803  199717
                             FI 97392         A   19970130
NO 9700476     A   19970203  WO 95EP3098      A   19950803  199718
                             NO 97476         A   19970203
JP 9508774     W   19970902  WO 95EP3098      A   19950803  199745
                             JP 96506218      A   19950803
BR 9508483     A   19971125  BR 958483        A   19950803  199803
                             WO 95EP3098      A   19950803
US 5930249     A   19990727  WO 95EP3098      A   19950803  199936
                             US 97776561      A   19970131
JP 3043421     B2  20000522  WO 95EP3098      A   19950803  200029
                             JP 96506218      A   19950803

Priority Applications (No Type Date): EP 94112147 A 19940803
Cited Patents: 04Jnl.Ref; EP 372270; EP 376556; EP 490446; EP 229494
Patent Details:
Patent No  Kind Lan Pg   Main IPC     Filing Notes
EP 696147     A1 G   8 H04Q-003/66
   Designated States (Regional): AT BE CH DE DK ES FR GB GR IE IT LI LU NL
   PT SE
JP 3043421    B2     6 H04M-003/00   Previous Publ. patent JP 9508774
```

[redacted] Based on patent WO [redacted]757

WO 9604757 A1 G 20 H04Q-003/66
Designated States (National): BR CN FI JP NO RU US
JP 9508774 W 21 H04M-003/00 Based on patent WO 9604757
BR 9508483 A H04Q-003/66 Based on patent WO 9604757
US 5930249 A H04J-001/16 Based on patent WO 9604757
FI 9700392 A H04Q-000/00
NO 9700476 A H04Q-000/00

Abstract (Basic): EP 696147 A

The traffic routing system has the calls placed between a source exchange (1) and a target exchange 92) placed via one or more defined paths, with an alternative path selected via a defined selection plan when no defined path is available, for handling the traffic overload.

The overload traffic is cyclically distributed among the alternative connection paths, with the call capacitance of each connection path held by the routing system. Pref. each newly established connection path is occupied by a given number of overload calls, before the next alternative connection path is established.

ADVANTAGE - Efficient handling of overload calls.

Dwg.2/2

Title Terms: DYNAMIC; TRAFFIC; ROUTE; SYSTEM; COMMUNICATE; NETWORK; ALTERNATIVE; CALL; CONNECT; PATH; ESTABLISH; DEFINE; CONNECT; PATH; SOURCE; TARGET; EXCHANGE; OCCUPY
Derwent Class: W01
International Patent Class (Main): H04J-001/16; H04M-003/00; H04Q-000/00; H04Q-003/66
International Patent Class (Additional): H04L-012/50; H04Q-003/545; H04Q-003/64
File Segment: EPI
?logoff

06jul00 10:55:28 User212334 Session D2254.2
Sub account: P001131
$4.27 0.194 DialUnits File351
$3.76 1 Type(s) in Format 5
$3.76 1 Types
$8.03 Estimated cost File351
$0.19 TYMNET
$8.22 Estimated cost this search
$8.30 Estimated total session cost 0.357 DialUnits